Manual Functional Annotation

Manual Functional Annotation

In this class, we will cover:

- The rationale behind manual annotation
- The process of annotating eukaryotic genes manually
- Software tools we use for manual annotation
- Steps you can take to annotate or verify an annotatiom

Uses for Annotation Knowledge

- Understanding and assessing quality of existing annotations
- Annotating a new genome
- Reannotating an existing genome

Evaluating existing annotations

A gene accession usually has information associated with it.

- How did it get its name?
- How plausible is the function assigned to it?
- Where did this information come from?
- Is the information accurate? Can you rely on it?

Goals of the Annotation Process

Some of the goals of annotation of gene products are:

- to determine the function of the protein, if possible;
- to assign attributes to the protein: functional name, symbol, GO terms, comments as needed;
- to be as specific as evidence supports, erring on the side of accuracy rather than specificity;
- to store supporting evidence for the assigned attributes;
- to make the information available as appropriate.

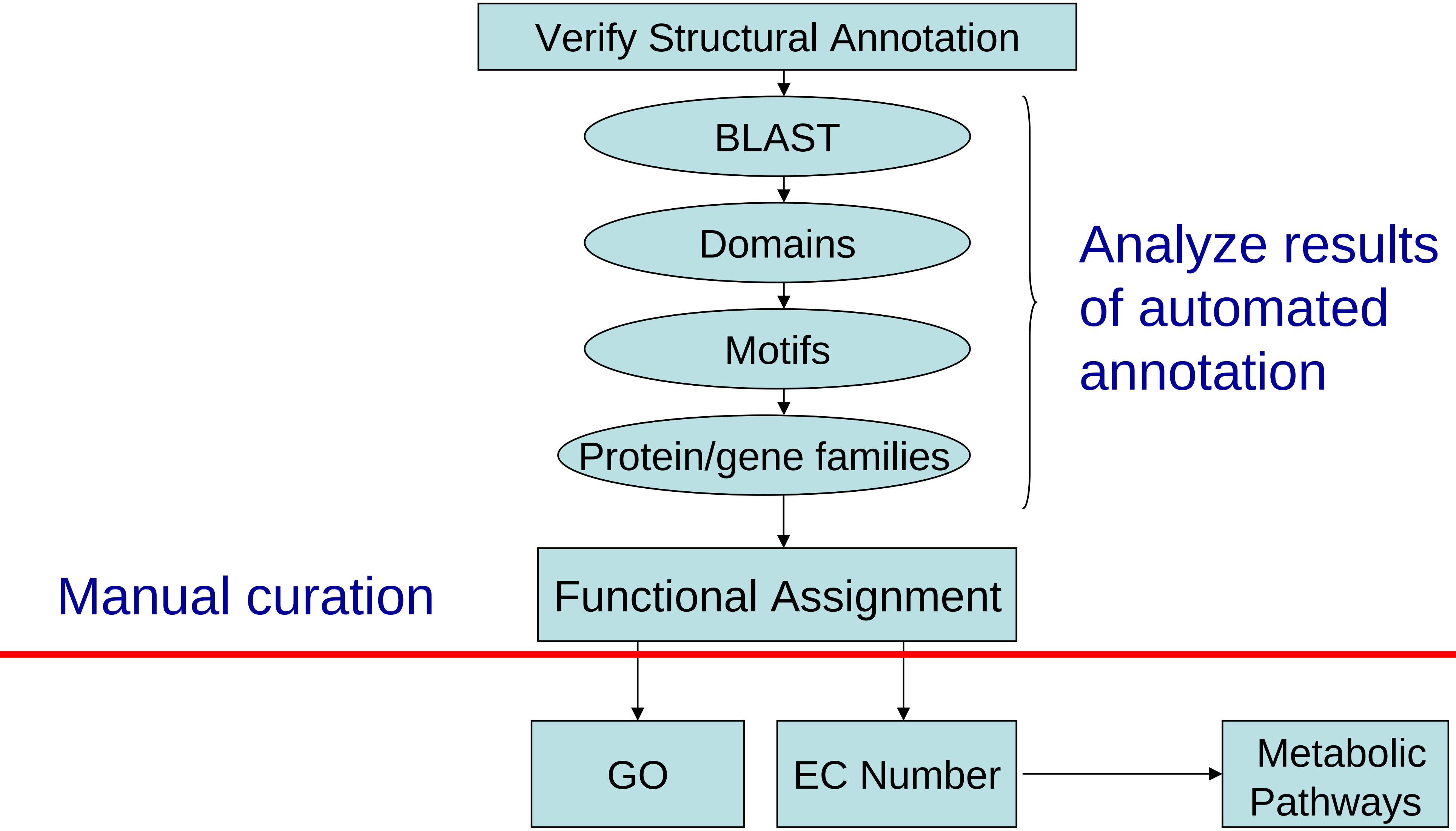
Manual vs. automated annotations

Automated annotation:

- derived from computational approaches
- use of different methods at different centers
- complicated by high volumes of data

The highest quality annotation often requires **manual** review and intervention.

Functional Annotation



Verify Structural Annotation

BLAST

Domains

Motifs

Protein/gene families

Analyze results of automated annotation

Functional Assignment

Manual curation

GO

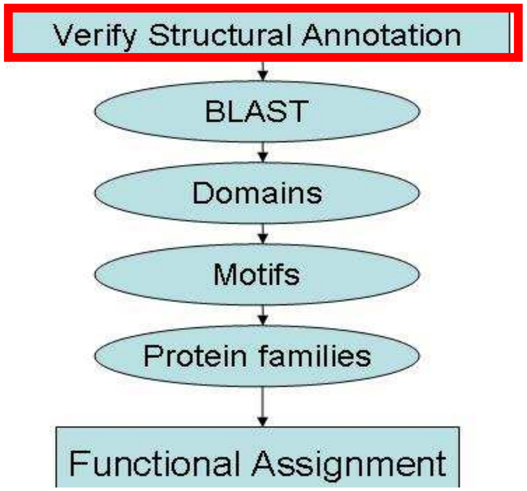
EC Number

Metabolic Pathways

First, verify the gene structure

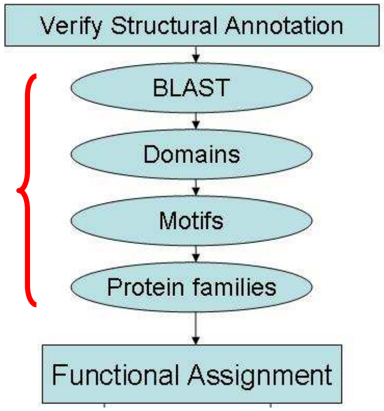
- Check to be sure the gene structure before you put effort into the functional annotation:
  - Look at the evidence
  - Verify against EST/cDNA, BLAST hits…
- Correct the gene structure

  if necessary.



Verify evidence from automated annotation

- BLAST matches
- Domains
- Prosite, Interpro classifications
- Motifs
- Signal Sequence
- Target Sequence
- EC number
- Transmembrane domain(s)
- Paralogous families



Homology Searching for Functional Annotation

Tools that are available to help you characterize a sequence

- **WU BLAST** http://blast.wustl.edu/ with links to many servers
- **NCBI BLAST** http://www.ncbi.nlm.nih.gov/blast/
- **Pfam profiles** (profiles, or HMMs) http://pfam.wustl.edu/
- **TIGRFAMS** (profiles, or HMMs) http://tigrblast.tigr.org/web-hmm/
- **SCOP** (profiles, or HMMs) http://iris.physics.iisc.ernet.in/scop/
- **CDD** (conserved domain database) http://www.ncbi.nlm.nih.gov/Structure/cdd/cdd.shtml
- **Prosite** (profiles & families) http://ca.expasy.org/tools/scanprosite/
- **Interpro** (families) http://www.ebi.ac.uk/InterProScan/
- **Swiss-Prot** **http://au.expasy.org/sprot/**
- **TmHMM** (transmembrane domain) http://www.cbs.dtu.dk/services/TMHMM/
- **SignalP** (signal peptide cleavage sites) http://www.cbs.dtu.dk/services/SignalP/
- **TargetP** (subcellular location) http://www.cbs.dtu.dk/services/TargetP/
- **PSI-BLAST** (NCBI) link at http://www.ncbi.nlm.nih.gov/BLAST/
- **Protein families and clustering**
  - **JCVI Paralogous Families** (not yet available outside of JCVI)
  - **TribeMCL** http://micans.org/mcl/
  - **Superfamily** http://supfam.mrc-lmb.cam.ac.uk/SUPERFAMILY/

Databases to search

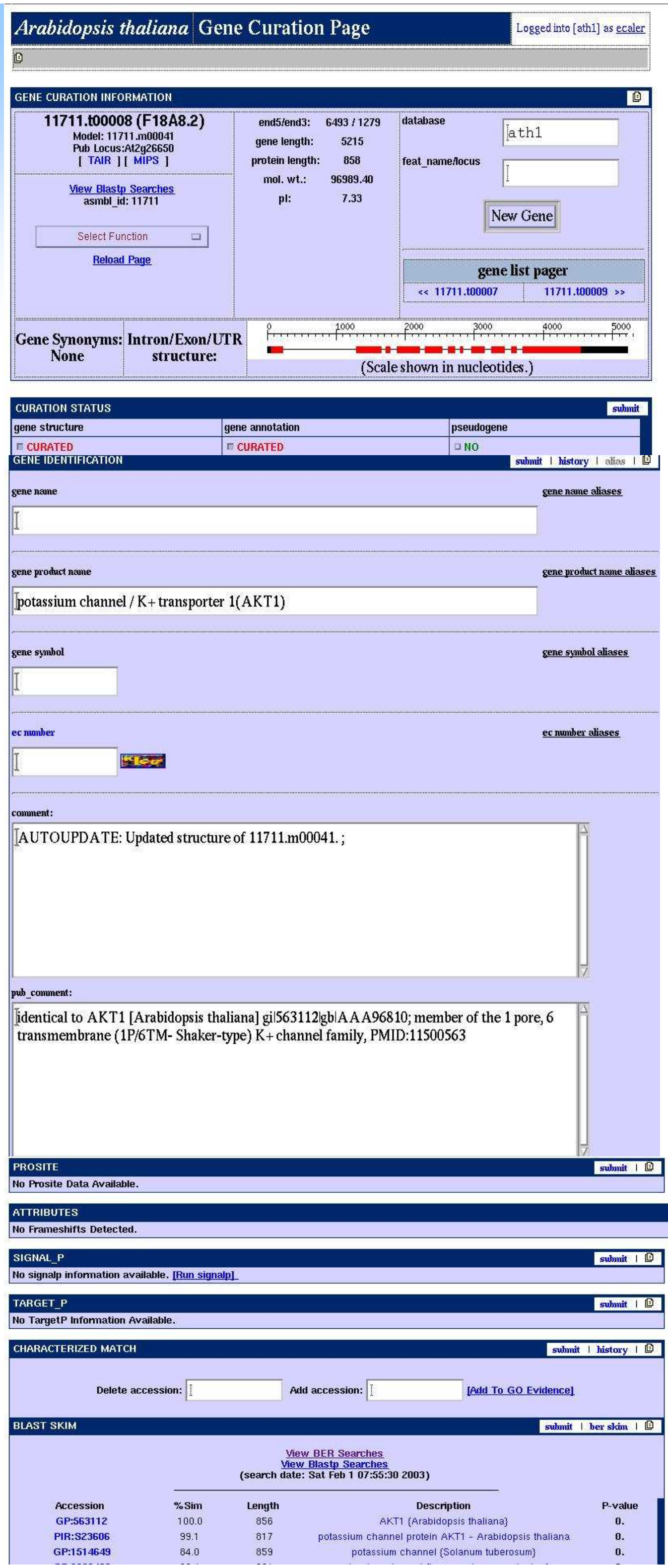
- NCBI Blast http://www.ncbi.nlm.nih.gov/blast/
- JCVI/TIGR eukaryotic databases http://www.tigr.org/tdb/euk/ (follow links to each database)
- JCVI/TIGR Blast (Rice, Arabidopsis) http://tigrblast.tigr.org/euk-blast/index.cgi?project=osa1
- Dana Farber Gene Indices http://compbio.dfci.harvard.edu/tgi/tgipage.html
- JCVI CMR (microbial) http://tigrblast.tigr.org/cmr-blast/
- Sanger projects http://www.sanger.ac.uk/DataSearch/
- WU GSC Blast Server http://genome.wustl.edu/tools/blast/

…and many others

Manatee

- Manatee is a web-based gene evaluation and genome annotation tool.
- Manatee can store and present annotation for prokaryotic and eukaryotic genomes.
- We use Manatee for manual annotation.
  You can, too, if you have the support of an IT department, or a capable engineer.
- Download it at:

http://sourceforge.net/projects/manatee/



Arabidopsis thaliana Gene Curation Page

Logged into [ath1] as ecaler

GENE CURATION INFORMATION

11711.t00008 (F18A8.2)
Model: 11711.m00041
Pub Locus:At2g26650
[TAIR] [MIPS]

View Blastp Searches
asmbl_id: 11711

end5/end3:	6493 / 1279
gene length:	5215
protein length:	858
mol. wt.:	96989.40
pI:	7.33

database: ath1

feat_name/locus:

gene list pager
<< 11711.t00007 | 11711.t00009 >>

Gene Synonyms: None

Intron/Exon/UTR structure:

(Scale shown in nucleotides.)

CURATION STATUS

gene structure	gene annotation	pseudogene
☒ CURATED	☒ CURATED	□ NO

GENE IDENTIFICATION

gene name:

gene product name: potassium channel / K+ transporter 1(AKT1)

gene symbol:

ec number:

comment: AUTOUPDATE: Updated structure of 11711.m00041. ;

pub_comment: identical to AKT1 [Arabidopsis thaliana] gi|563112|gb|AAA96810; member of the 1 pore, 6 transmembrane (1P/6TM- Shaker-type) K+ channel family, PMID:11500563

PROSITE

No Prosite Data Available.

ATTRIBUTES

No Frameshifts Detected.

SIGNAL_P

No signalp information available. [Run signalp]

TARGET_P

No TargetP Information Available.

CHARACTERIZED MATCH

Delete accession: Add accession: [Add To GO Evidence]

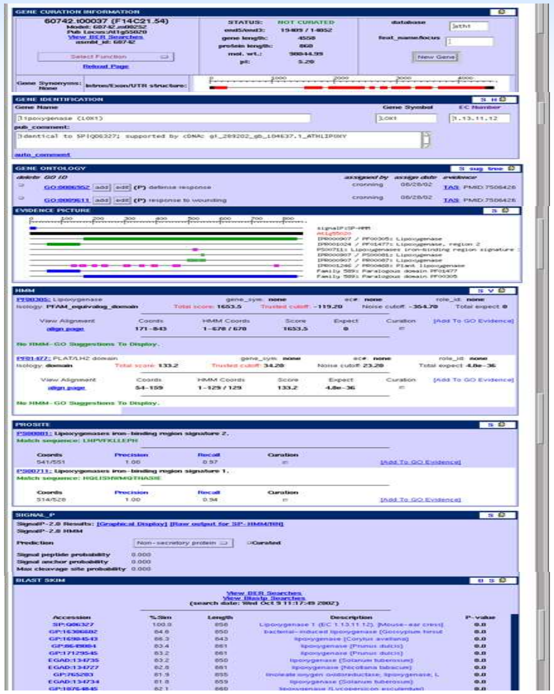
BLAST SKIM

View BER Searches
View Blastp Searches
(search date: Sat Feb 1 07:55:30 2003)

Accession	%Sim	Length	Description	P-value
GP:563112	100.0	856	AKT1 {Arabidopsis thaliana}	0.
PIR:S23606	99.1	817	potassium channel protein AKT1 - Arabidopsis thaliana	0.
GP:1514649	84.0	859	potassium channel {Solanum tuberosum}	0.

Use all possible resources…

Pfam
Protein families database of alignments and HMMs
Wellcome Trust Sanger Institute
Chalcone and stilbene synthases, C-terminal domain
INTERPRO description
FUNCTION
PROCESS


GENE CURATION INFORMATION

60742.t00037 (F14C21.54)
Model: 60742.m00252
Pub Locus:At1g55020
View BER Searches
asmbl_id: 60742

STATUS: NOT CURATED
end5/end3: 19409 / 14852
gene length: 4558
protein length: 860
mol. wt.: 98844.99
pI: 5.20

database: atht
feat_name/focus:
New Gene

Select Function
Reload Page

Gene Synonyms: None
Intron/Exon/UTR structure:

GENE IDENTIFICATION

Gene Name: lipoxygenase (LOX1)
Gene Symbol: LOX1
EC Number: 1.13.11.12
pub_comment: identical to SP|Q06327; supported by cDNA: gi_289202_gb_L04637.1_ATHLIPOXY
auto_comment

GENE ONTOLOGY

delete GO ID / assigned by / assign date / evidence
GO:0006952 add edit (P) defense response — cronning 08/28/02 TAS PMID:7506426
GO:0009611 add edit (P) response to wounding — cronning 08/28/02 TAS PMID:7506426

EVIDENCE PICTURE

signalP:SP-HMM
At1g55020
IPR000907 / PF00305: Lipoxygenase
IPR001024 / PF01477: Lipoxygenase, region 2
PS00711: Lipoxygenases iron-binding region signature 1
IPR000907 / PS00081: Lipoxygenase
IPR000907 / PR00087: Lipoxygenase
IPR001246 / PR00468: Plant lipoxygenase
Family 589: Paralogous domain PF01477
Family 589: Paralogous domain PF00305

HMM

PF00305: Lipoxygenase — gene_sym: none — ec#: none — role_id: none
Isology: PFAM_equivalog_domain — Total score: 1653.5 — Trusted cutoff: -119.20 — Noise cutoff: -354.70 — Total expect: 0

View Alignment	Coords	HMM Coords	Score	Expect	Curation	[Add To GO Evidence]
align page	171-843	1-678 / 678	1653.5	0		

No HMM-GO Suggestions To Display.

PF01477: PLAT/LH2 domain — gene_sym: none — ec#: none — role_id: none
Isology: domain — Total score: 133.2 — Trusted cutoff: 34.20 — Noise cutoff: 23.20 — Total expect: 4.8e-36

View Alignment	Coords	HMM Coords	Score	Expect	Curation	[Add To GO Evidence]
align page	54-159	1-129 / 129	133.2	4.8e-36		

No HMM-GO Suggestions To Display.

PROSITE

PS00081: Lipoxygenases iron-binding region signature 2.
Match sequence: LHPVFKLLEPH

Coords	Precision	Recall	Curation	
541/551	1.00	0.97		[Add To GO Evidence]

PS00711: Lipoxygenases iron-binding region signature 1.
Match sequence: HQLISHWLRTHASIE

Coords	Precision	Recall	Curation	
514/528	1.00	0.94		[Add To GO Evidence]

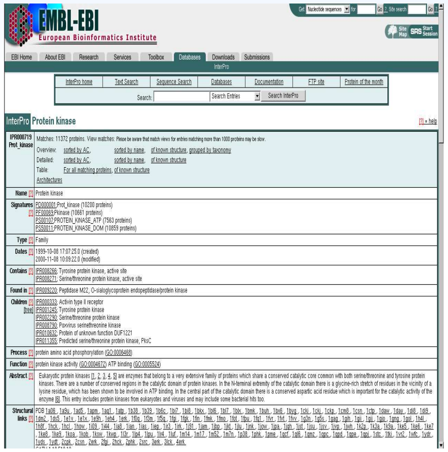
SIGNAL_P

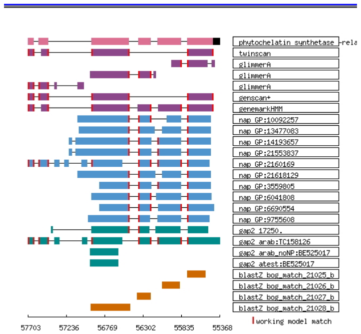
SignalP-2.0 Results: [Graphical Display] [Raw output for SP-HMM/NN]
SignalP-2.0 HMM
Prediction: Non-secretory protein □Curated
Signal peptide probability 0.000
Signal anchor probability 0.000
Max cleavage site probability 0.000

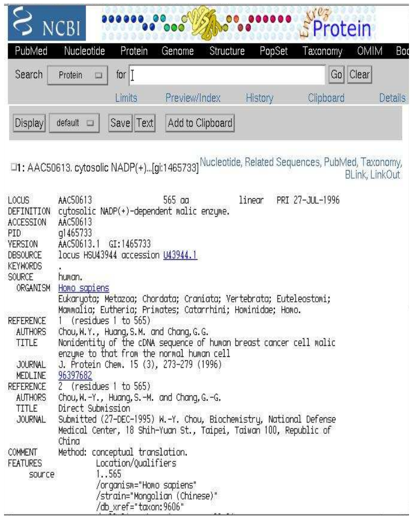
BLAST SKIM

View BER Searches
View Blastp Searches
(search date: Wed Oct 9 11:17:49 2002)

Accession	%Sim	Length	Description	P-value
SP:Q06327	100.0	856	Lipoxygenase 1 (EC 1.13.11.12) (Mouse-ear cress)	0.0
GP:16306682	84.6	850	bacterial-induced lipoxygenase (Gossypium hirsut	0.0
GP:16904543	86.3	843	lipoxygenase (Corylus avellana)	0.0
GP:8649084	83.4	861	lipoxygenase (Prunus dulcis)	0.0
GP:17129545	83.2	861	lipoxygenase (Prunus dulcis)	0.0
EGAD:134735	83.2	850	lipoxygenase (Solanum tuberosum)	0.0
EGAD:134727	82.8	861	lipoxygenase (Nicotiana tabacum)	0.0
GP:765283	81.9	855	linoleate oxygen oxidoreductase; lipoxygenase; L	0.0
EGAD:134734	81.8	859	lipoxygenase (Solanum tuberosum)	0.0
GP:1676845	82.1	860	lipoxygenase (Lycopersicon esculentum)	0.0



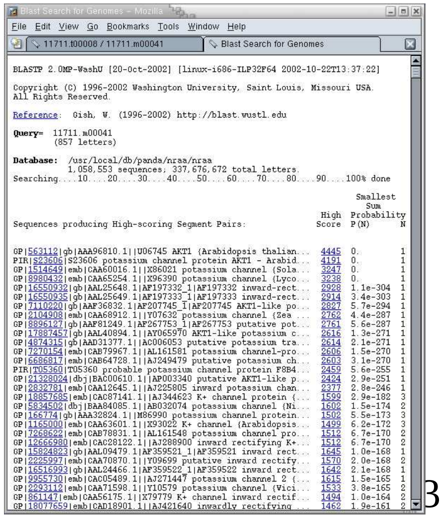



NCBI
Entrez Protein
PubMed Nucleotide Protein Genome Structure PopSet Taxonomy OMIM Boo
Search Protein for
Go Clear
Limits Preview/Index History Clipboard Details
Display default Save Text Add to Clipboard

□1: AAC50613. cytosolic NADP(+)...[gi:1465733] Nucleotide, Related Sequences, PubMed, Taxonomy, BLink, LinkOut

```
LOCUS       AAC50613                 565 aa            linear   PRI 27-JUL-1996
DEFINITION  cytosolic NADP(+)-dependent malic enzyme.
ACCESSION   AAC50613
PID         g1465733
VERSION     AAC50613.1  GI:1465733
DBSOURCE    locus HSU43944 accession U43944.1
KEYWORDS    .
SOURCE      human.
  ORGANISM  Homo sapiens
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Mammalia; Eutheria; Primates; Catarrhini; Hominidae; Homo.
REFERENCE   1  (residues 1 to 565)
  AUTHORS   Chou,W.Y., Huang,S.M. and Chang,G.G.
  TITLE     Nonidentity of the cDNA sequence of human breast cancer cell malic
            enzyme to that from the normal human cell
  JOURNAL   J. Protein Chem. 15 (3), 273-279 (1996)
  MEDLINE   96397682
REFERENCE   2  (residues 1 to 565)
  AUTHORS   Chou,W.-Y., Huang,S.-M. and Chang,G.-G.
  TITLE     Direct Submission
  JOURNAL   Submitted (27-DEC-1995) W.-Y. Chou, Biochemistry, National Defense
            Medical Center, 18 Shih-Yuan St., Taipei, Taiwan 100, Republic of
            China
COMMENT     Method: conceptual translation.
FEATURES             Location/Qualifiers
     source          1..565
                     /organism="Homo sapiens"
                     /strain="Mongolian (Chinese)"
                     /db_xref="taxon:9606"
```

NCBI
PubMed
National Library of Medicine NLM
PubMed Nucleotide Protein Genome Structure PMC Taxonomy OMIM Books
Search PubMed for cyclic nucleotide arabidopsis Go Clear
Limits Preview/Index History Clipboard Details
Display Summary Show 20 Sort Send to Text
Page 1 of 3 Next
THE PLANT CELL
nature
Plasmodium genome
The EMBO Journal
nature biotechnology
the plant journal
The Journal of Biological Chemistry
PNAS
Measuring force at cell adhesions
nature genetics
Cell
Science
Folding of LDL Receptor Family Proteins


Blast Search for Genomes - Mozilla
File Edit View Go Bookmarks Tools Window Help
11711.t00008 / 11711.m00041
Blast Search for Genomes

```
BLASTP 2.0MP-WashU [20-Oct-2002] [linux-i686-ILP32F64 2002-10-22T13:37:22]

Copyright (C) 1996-2002 Washington University, Saint Louis, Missouri USA.
All Rights Reserved.

Reference:  Gish, W. (1996-2002) http://blast.wustl.edu

Query=  11711.m00041
        (857 letters)

Database:  /usr/local/db/panda/nraa/nraa
           1,058,553 sequences; 337,676,672 total letters.
Searching....10....20....30....40....50....60....70....80....90....100% done

                                                                    Smallest
                                                                      Sum
                                                             High  Probability
Sequences producing High-scoring Segment Pairs:              Score  P(N)      N

GP|563112|gb|AAA96810.1||U06745 AKT1 (Arabidopsis thalian...  4445  0.         1
PIR|S23606|S23606 potassium channel protein AKT1 - Arabid...  4191  0.         1
GP|1514649|emb|CAA60016.1||X86021 potassium channel (Sola...  3247  0.         1
GP|8980432|emb|CAA65254.1||X96390 potassium channel (Lyco...  3238  0.         1
GP|16550932|gb|AAL25648.1|AF197332_1|AF197332 inward-rect...  2928  1.1e-304  1
GP|16550935|gb|AAL25649.1|AF197333_1|AF197333 inward-rect...  2914  3.4e-303  1
GP|7110220|gb|AAF36832.1|AF207745_1|AF207745 AKT1-like po...  2827  5.7e-294  1
GP|2104908|emb|CAA68912.1||Y07632 potassium channel (Zea ...  2762  4.4e-287  1
GP|8896127|gb|AAF81249.1|AF267753_1|AF267753 putative pot...  2761  5.6e-287  1
GP|17887457|gb|AAL40894.1||AY065970 AKT1-like potassium c...  2616  1.3e-271  1
GP|4874315|gb|AAD31377.1||AC006053 putative potassium tra...  2614  2.1e-271  1
GP|7270154|emb|CAB79967.1||AL161581 potassium channel-pro...  2606  1.5e-270  1
GP|6686817|emb|CAB64728.1||AJ249479 putative potassium ch...  2603  3.1e-270  1
PIR|T05360|T05360 probable potassium channel protein F8B4...  2459  5.6e-255  1
GP|21328024|dbj|BAC00610.1||AP003340 putative AKT1-like p...  2424  2.9e-251  1
GP|2832781|emb|CAA12645.1||AJ225805 inward potassium chan...  2377  2.8e-246  1
GP|18857685|emb|CAC87141.1||AJ344623 K+ channel protein (...  1599  2.9e-182  3
GP|5834502|dbj|BAA84085.1||AB032074 potassium channel (Ni...  1602  1.5e-174  2
GP|166774|gb|AAA32824.1||M86990 potassium channel protein...  1502  5.5e-173  3
GP|1165000|emb|CAA63601.1||X93022 K+ channel (Arabidopsis...  1499  6.2e-172  3
GP|7268622|emb|CAB78831.1||AL161548 potassium channel pro...  1512  6.7e-170  2
GP|12666980|emb|CAC28122.1||AJ288900 inward rectifying K+...  1512  6.7e-170  2
GP|15824823|gb|AAL09479.1|AF359521_1|AF359521 inward rect...  1645  1.0e-168  1
GP|2225997|emb|CAA70870.1||Y09699 putative inward rectify...  1570  2.0e-168  2
GP|16516993|gb|AAL24466.1|AF359522_1|AF359522 inward rect...  1642  2.1e-168  1
GP|9955730|emb|CAC05489.1||AJ271447 potassium channel 2 (...  1615  1.5e-165  1
GP|2293112|emb|CAA71598.1||Y10579 potassium channel (Vici...  1533  3.8e-165  2
GP|861147|emb|CAA56175.1||X79779 K+ channel inward rectif...  1494  1.0e-164  2
GP|18077659|emb|CAD18901.1||AJ421640 inwardly rectifying ...  1462  1.9e-161  2
```

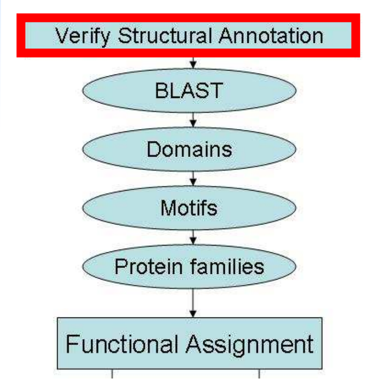
Example 1

Our first example will be a protein sequence from *Trypanosoma brucei*. Our task will be to annotate this protein sequence as fully as possible, given the tools at hand.

protein sequence:

```
>unknown_T. brucei protein_sequence
MLRRLGVRHFRRTPLLFVGGDGSIFERYTE
IDNSNERRINALKGCGMFEDEWIATEKVHG
ANFGIYSIEGEKMIRYAKRSGIMPPNEHFFG
YHILIPELQRYITSIREMLCEKQKKKLHVVL
INGELFGGKYDHPSVPKTRKTVMVAGKPR
TISAVQTDSFPQYSPDLHFYAFDIKYKETED
GDYTTLVYDEAIELFQRVPGLLYARAVIRG
PMSKVAAFDVERFVTTIPPLVGMGNYPLTG
NWAEGLVVKHSRLGMAGFDPKGPTVLKF
KCTAFQEISTDRAQGPRVDEMRNVRRDSIN
RAGVQLPDLESIVQDPIQLEASKLLLNHVCE
NRLKNVLSKIGTEPFEKEEMTPDQLATLLA
KDVLKDFLKDTEPSIVNIPVLIRKDLTRYVIF
ESRRLVCSQWKDILKRQSPDFSE*
```

Verify the gene structure



Verify Structural Annotation

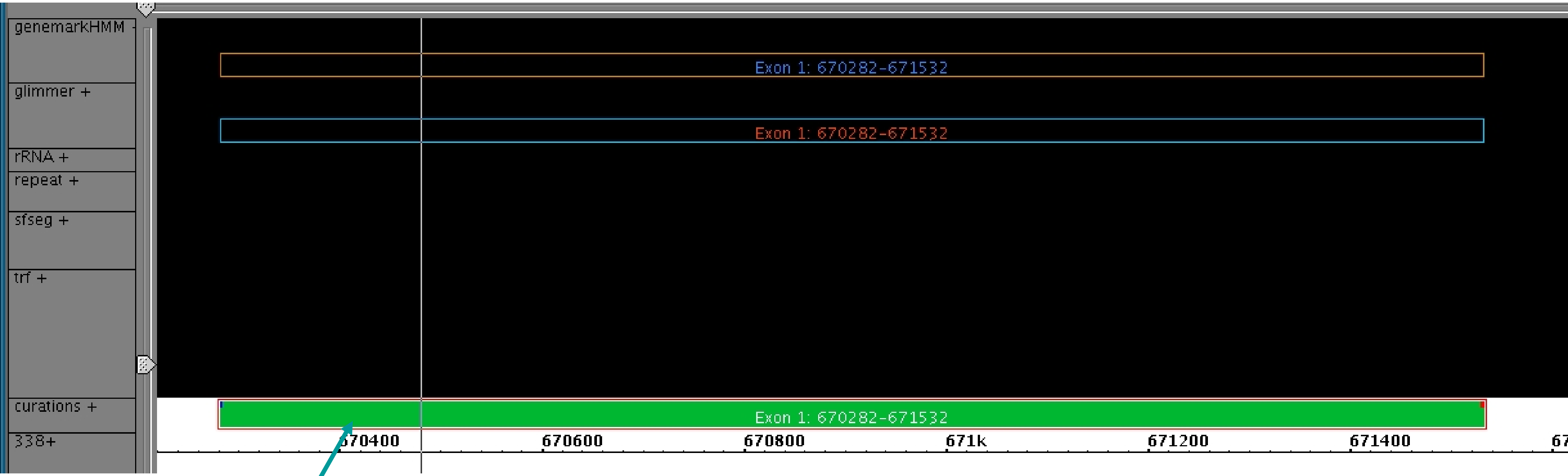
BLAST

Domains

Motifs

Protein families

Functional Assignment

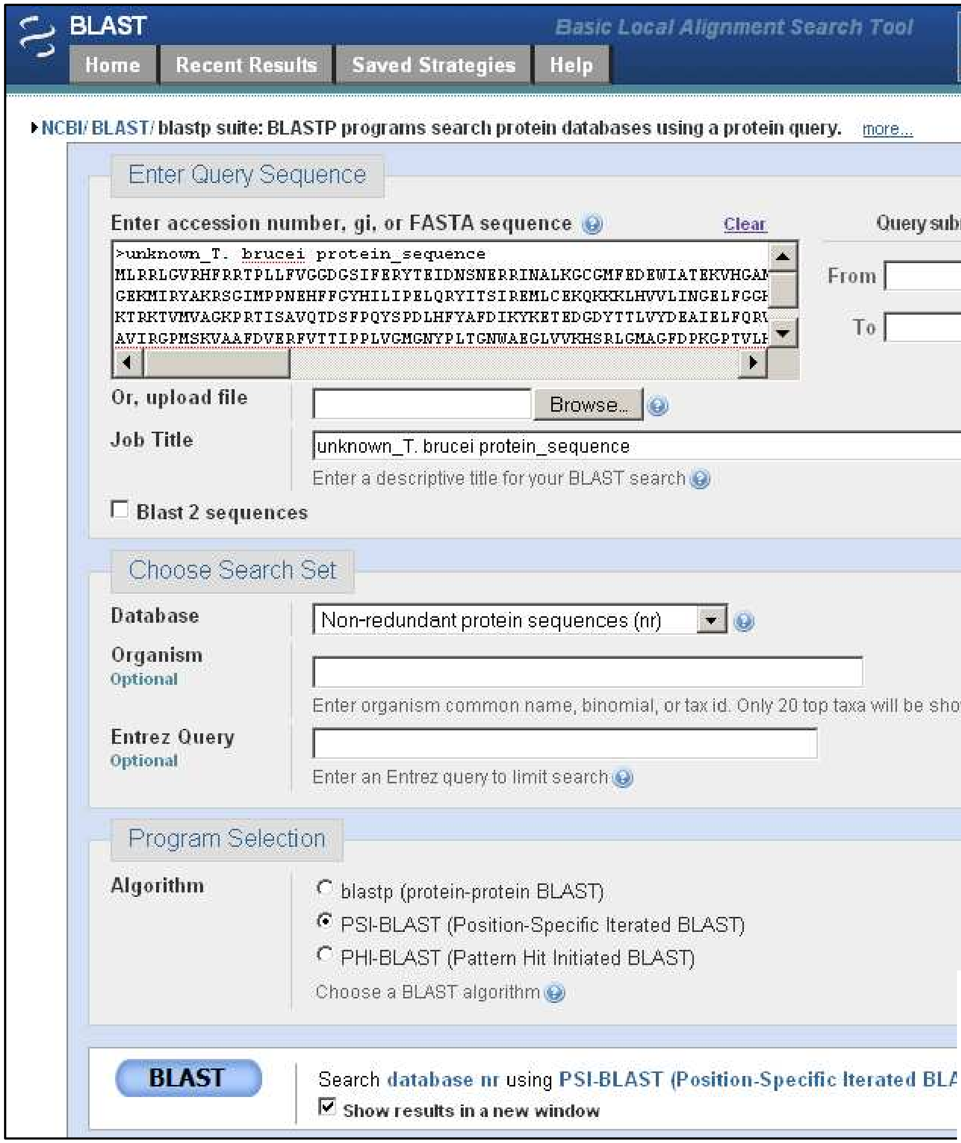


Gene

NCBI BLAST

NCBI BLAST tools at:
http://www.ncbi.nlm.nih.gov/blast/.

Program	Database	Query
BLASTN	Nucleotide	Nucleotide
BLASTP	Protein	Protein
BLASTX	Protein	Nucleotide → Protein
TBLASTN	Nucleotide →Protein	Protein
TBLASTX	Nucleotide →Protein	Nucleotide →Protein


BLAST
Basic Local Alignment Search Tool
Home
Recent Results
Saved Strategies
Help
My NCBI
Welcome lhannick. [Sign Out]

▸NCBI/ BLAST/ blastp suite: BLASTP programs search protein databases using a protein query. more...

Reset page

Enter Query Sequence

Enter accession number, gi, or FASTA sequence

Clear

```
>unknown_T. brucei protein_sequence
MLRRLGVRHFRRTPLLFVGGDGSIFERYTEIDNSNERRINALKGCGMFEDEWIATEKVHGA
GEKMIRYAKRSGIMPPNEHFFGYHILIPELQRYITSIREMLCEKQKKKLHVVLINGELFGG
KTRKTVMVAGKPRTISAVQTDSFPQYSPDLHFYAFDIKYKETEDGDYTTLVYDEAIELFQR
AVIRGPMSKVAAFDVERFVTTIPPLVGMGNYPLTGNWAEGLVVKHSRLGMAGFDPKGPTVL
```

Query subrange

From

To

Or, upload file Browse...

Job Title: unknown_T. brucei protein_sequence

Enter a descriptive title for your BLAST search

□ Blast 2 sequences

Choose Search Set

Database: Non-redundant protein sequences (nr)

Organism (Optional):

Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown.

Entrez Query (Optional):

Enter an Entrez query to limit search

Program Selection

Algorithm

○ blastp (protein-protein BLAST)

● PSI-BLAST (Position-Specific Iterated BLAST)

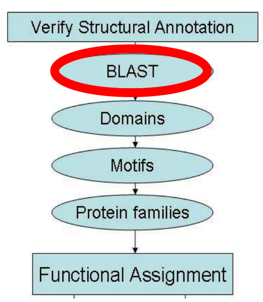
○ PHI-BLAST (Pattern Hit Initiated BLAST)

Choose a BLAST algorithm

BLAST

Search database nr using PSI-BLAST (Position-Specific Iterated BLA

☑ Show results in a new window



Read → as "translated to"

BLAST: What makes a good alignment?

It depends on what you are trying to prove!

- minimum of 35% identity, better 40% & up
  - higher for short proteins
  - score is weighted for length

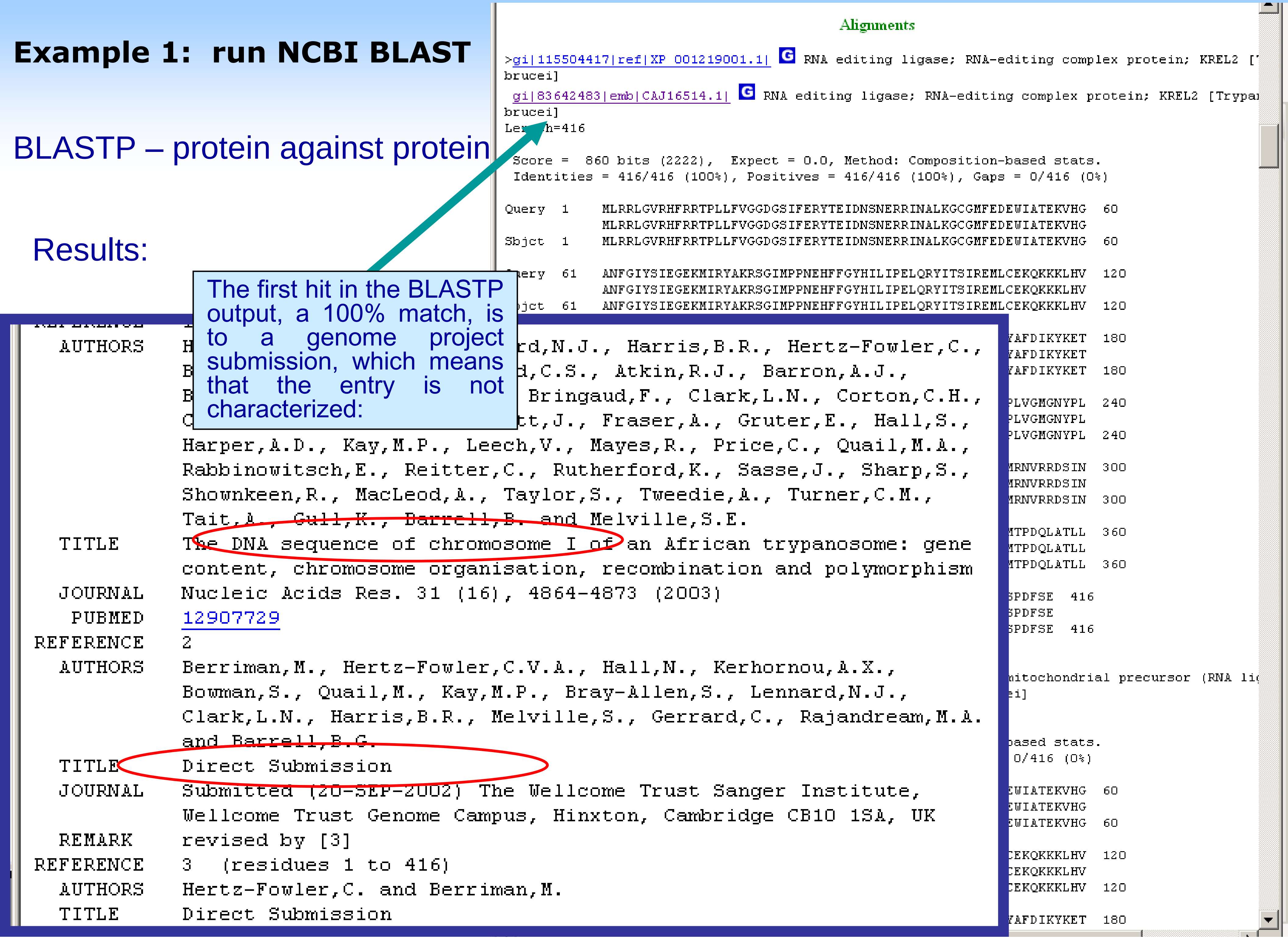
- full length match
  - at least 80% of both proteins

? See explanation of BLAST scores in extra slides.

Example 1: run NCBI BLAST

BLASTP – protein against protein

Results:



The first hit in the BLASTP output, a 100% match, is to a genome project submission, which means that the entry is not characterized:

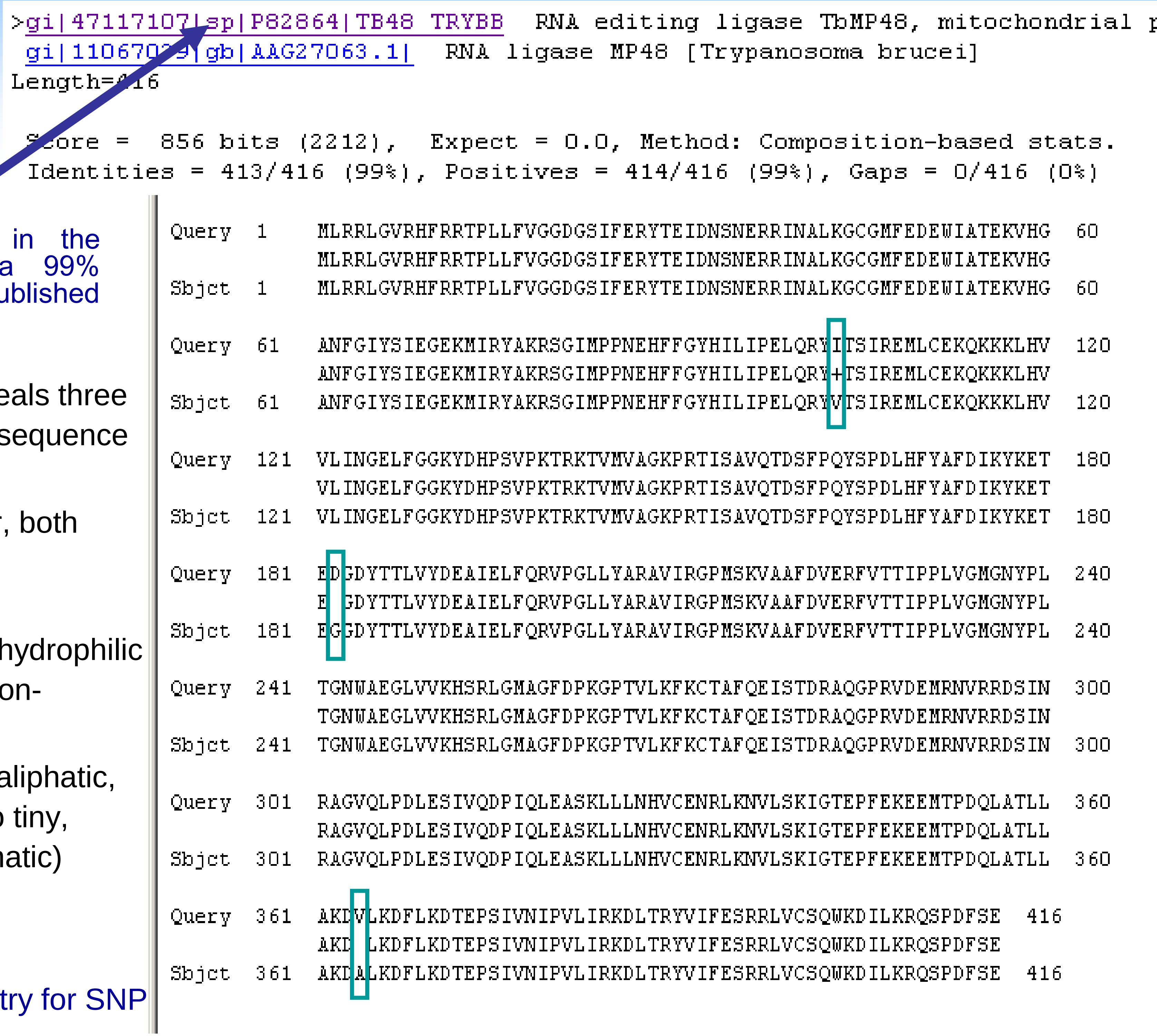
Alignments

```
>gi|115504417|ref|XP_001219001.1| G RNA editing ligase; RNA-editing complex protein; KREL2 [T
brucei]
 gi|83642483|emb|CAJ16514.1| G RNA editing ligase; RNA-editing complex protein; KREL2 [Trypa
brucei]
Length=416

 Score =  860 bits (2222),  Expect = 0.0, Method: Composition-based stats.
 Identities = 416/416 (100%), Positives = 416/416 (100%), Gaps = 0/416 (0%)

Query  1     MLRRLGVRHFRRTPLLFVGGDGSIFERYTEIDNSNERRINALKGCGMFEDEWIATEKVHG  60
             MLRRLGVRHFRRTPLLFVGGDGSIFERYTEIDNSNERRINALKGCGMFEDEWIATEKVHG
Sbjct  1     MLRRLGVRHFRRTPLLFVGGDGSIFERYTEIDNSNERRINALKGCGMFEDEWIATEKVHG  60

 ery  61    ANFGIYSIEGEKMIRYAKRSGIMPPNEHFFGYHILIPELQRYITSIREMLCEKQKKKLHV  120
             ANFGIYSIEGEKMIRYAKRSGIMPPNEHFFGYHILIPELQRYITSIREMLCEKQKKKLHV
 jct  61    ANFGIYSIEGEKMIRYAKRSGIMPPNEHFFGYHILIPELQRYITSIREMLCEKQKKKLHV  120

                                                              YAFDIKYKET  180
                                                              YAFDIKYKET
                                                              YAFDIKYKET  180

                                                              PLVGMGNYPL  240
                                                              PLVGMGNYPL
                                                              PLVGMGNYPL  240

                                                              MRNVRRDSIN  300
                                                              MRNVRRDSIN
                                                              MRNVRRDSIN  300

                                                              MTPDQLATLL  360
                                                              MTPDQLATLL
                                                              MTPDQLATLL  360

                                                              SPDFSE  416
                                                              SPDFSE
                                                              SPDFSE  416

                                                              mitochondrial precursor (RNA li
                                                              ei]

                                                              based stats.
                                                               0/416 (0%)

                                                              EWIATEKVHG  60
                                                              EWIATEKVHG
                                                              EWIATEKVHG  60

                                                              CEKQKKKLHV  120
                                                              CEKQKKKLHV
                                                              CEKQKKKLHV  120

                                                              YAFDIKYKET  180
```

```
  AUTHORS   H                      rd,N.J., Harris,B.R., Hertz-Fowler,C.,
            B                      d,C.S., Atkin,R.J., Barron,A.J.,
            B                      Bringaud,F., Clark,L.N., Corton,C.H.,
            C                      tt,J., Fraser,A., Gruter,E., Hall,S.,
            Harper,A.D., Kay,M.P., Leech,V., Mayes,R., Price,C., Quail,M.A.,
            Rabbinowitsch,E., Reitter,C., Rutherford,K., Sasse,J., Sharp,S.,
            Shownkeen,R., MacLeod,A., Taylor,S., Tweedie,A., Turner,C.M.,
            Tait,A., Gull,K., Barrell,B. and Melville,S.E.
  TITLE     The DNA sequence of chromosome I of an African trypanosome: gene
            content, chromosome organisation, recombination and polymorphism
  JOURNAL   Nucleic Acids Res. 31 (16), 4864-4873 (2003)
   PUBMED   12907729
REFERENCE   2
  AUTHORS   Berriman,M., Hertz-Fowler,C.V.A., Hall,N., Kerhornou,A.X.,
            Bowman,S., Quail,M., Kay,M.P., Bray-Allen,S., Lennard,N.J.,
            Clark,L.N., Harris,B.R., Melville,S., Gerrard,C., Rajandream,M.A.
            and Barrell,B.G.
  TITLE     Direct Submission
  JOURNAL   Submitted (20-SEP-2002) The Wellcome Trust Sanger Institute,
            Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SA, UK
  REMARK    revised by [3]
REFERENCE   3  (residues 1 to 416)
  AUTHORS   Hertz-Fowler,C. and Berriman,M.
  TITLE     Direct Submission
```

Example 1: navigating BLAST output

The second hit in the BLAST output, a 99% match, is to a published Swiss-Prot entry.



```
>gi|47117107|sp|P82864|TB48_TRYBB  RNA editing ligase TbMP48, mitochondrial p
 gi|11067029|gb|AAG27063.1|  RNA ligase MP48 [Trypanosoma brucei]
Length=416

 Score =  856 bits (2212),  Expect = 0.0, Method: Composition-based stats.
 Identities = 413/416 (99%), Positives = 414/416 (99%), Gaps = 0/416 (0%)

Query  1    MLRRLGVRHFRRTPLLFVGGDGSIFERYTEIDNSNERRINALKGCGMFEDEWIATEKVHG  60
            MLRRLGVRHFRRTPLLFVGGDGSIFERYTEIDNSNERRINALKGCGMFEDEWIATEKVHG
Sbjct  1    MLRRLGVRHFRRTPLLFVGGDGSIFERYTEIDNSNERRINALKGCGMFEDEWIATEKVHG  60

Query  61   ANFGIYSIEGEKMIRYAKRSGIMPPNEHFFGYHILIPELQRYITSIREMLCEKQKKKLHV  120
            ANFGIYSIEGEKMIRYAKRSGIMPPNEHFFGYHILIPELQRY+TSIREMLCEKQKKKLHV
Sbjct  61   ANFGIYSIEGEKMIRYAKRSGIMPPNEHFFGYHILIPELQRYVTSIREMLCEKQKKKLHV  120

Query  121  VLINGELFGGKYDHPSVPKTRKTVMVAGKPRTISAVQTDSFPQYSPDLHFYAFDIKYKET  180
            VLINGELFGGKYDHPSVPKTRKTVMVAGKPRTISAVQTDSFPQYSPDLHFYAFDIKYKET
Sbjct  121  VLINGELFGGKYDHPSVPKTRKTVMVAGKPRTISAVQTDSFPQYSPDLHFYAFDIKYKET  180

Query  181  EDGDYTTLVYDEAIELFQRVPGLLYARAVIRGPMSKVAAFDVERFVTTIPPLVGMGNYPL  240
            E GDYTTLVYDEAIELFQRVPGLLYARAVIRGPMSKVAAFDVERFVTTIPPLVGMGNYPL
Sbjct  181  EGGDYTTLVYDEAIELFQRVPGLLYARAVIRGPMSKVAAFDVERFVTTIPPLVGMGNYPL  240

Query  241  TGNWAEGLVVKHSRLGMAGFDPKGPTVLKFKCTAFQEISTDRAQGPRVDEMRNVRRDSIN  300
            TGNWAEGLVVKHSRLGMAGFDPKGPTVLKFKCTAFQEISTDRAQGPRVDEMRNVRRDSIN
Sbjct  241  TGNWAEGLVVKHSRLGMAGFDPKGPTVLKFKCTAFQEISTDRAQGPRVDEMRNVRRDSIN  300

Query  301  RAGVQLPDLESIVQDPIQLEASKLLLNHVCENRLKNVLSKIGTEPFEKEEMTPDQLATLL  360
            RAGVQLPDLESIVQDPIQLEASKLLLNHVCENRLKNVLSKIGTEPFEKEEMTPDQLATLL
Sbjct  301  RAGVQLPDLESIVQDPIQLEASKLLLNHVCENRLKNVLSKIGTEPFEKEEMTPDQLATLL  360

Query  361  AKDVLKDFLKDTEPSIVNIPVLIRKDLTRYVIFESRRLVCSQWKDILKRQSPDFSE  416
            AKD LKDFLKDTEPSIVNIPVLIRKDLTRYVIFESRRLVCSQWKDILKRQSPDFSE
Sbjct  361  AKDALKDFLKDTEPSIVNIPVLIRKDLTRYVIFESRRLVCSQWKDILKRQSPDFSE  416
```

The alignment reveals three positions with sequence variations:

I103V (very similar, both hydrophobic) conservative

D182G (negative, hydrophilic to tiny polar) non-conservative

V364A (nonpolar, aliphatic, hydrophobic to tiny, nonpolar, aliphatic) conservative

? See Glossary entry for SNP

Identity vs. similarity

- Identity means amino acids match exactly
- Similarity means the amino acids share either similar structure or properties (aromatic, hydrophilic, acidic, basic, etc) and thus MIGHT carry out the same or similar roles in the protein.

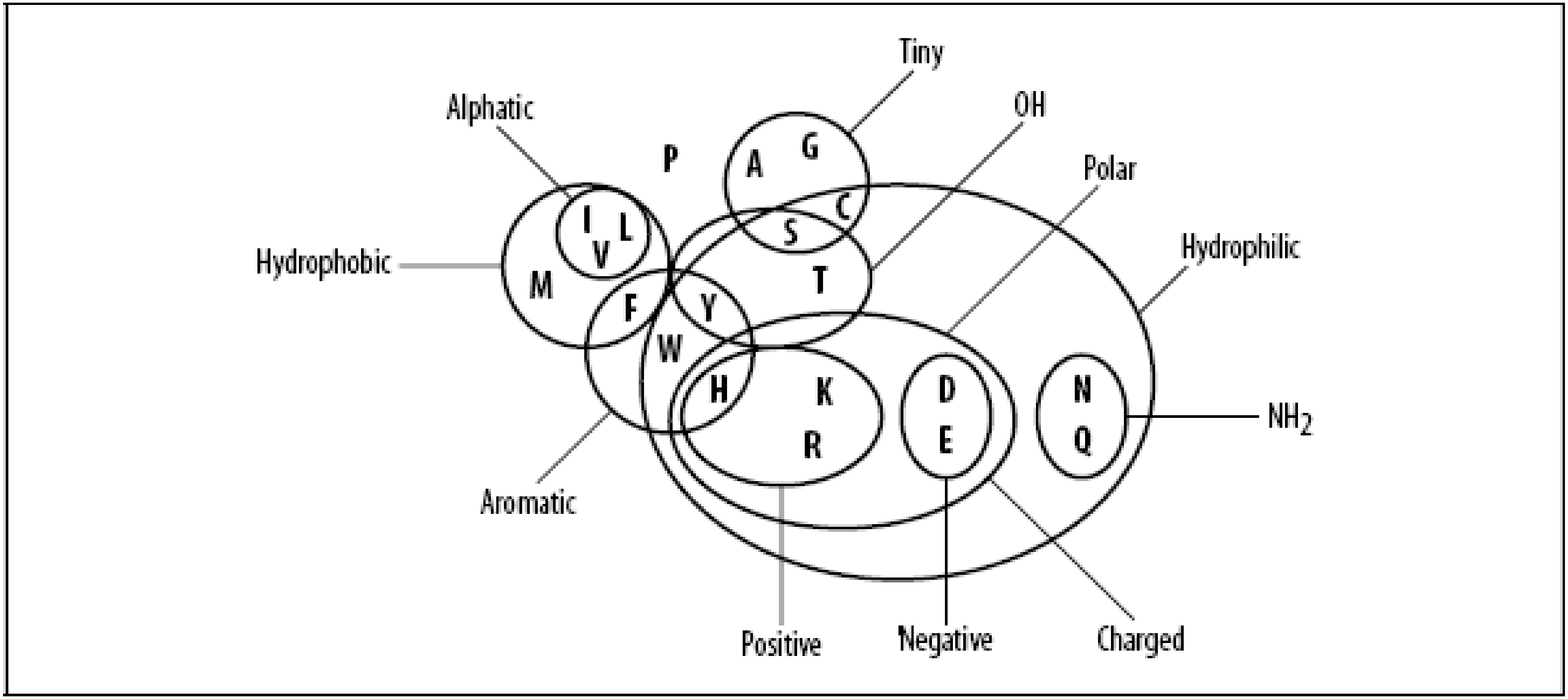


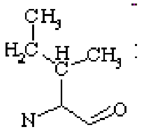
Figure 4-1. Amino acid chemical relationships

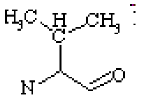
From O'Reilly BLAST (2003), Ch. 4

Differences in the amino acids

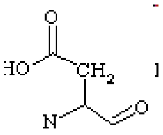
The alignment reveals three positions with sequence variations:

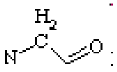
- I103V (very similar, both hydrophobic) conservative


[Ile]

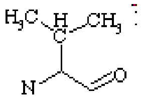

[Val]

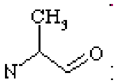
- D182G (negative, hydrophilic to tiny polar) non-conservative


[Asp]

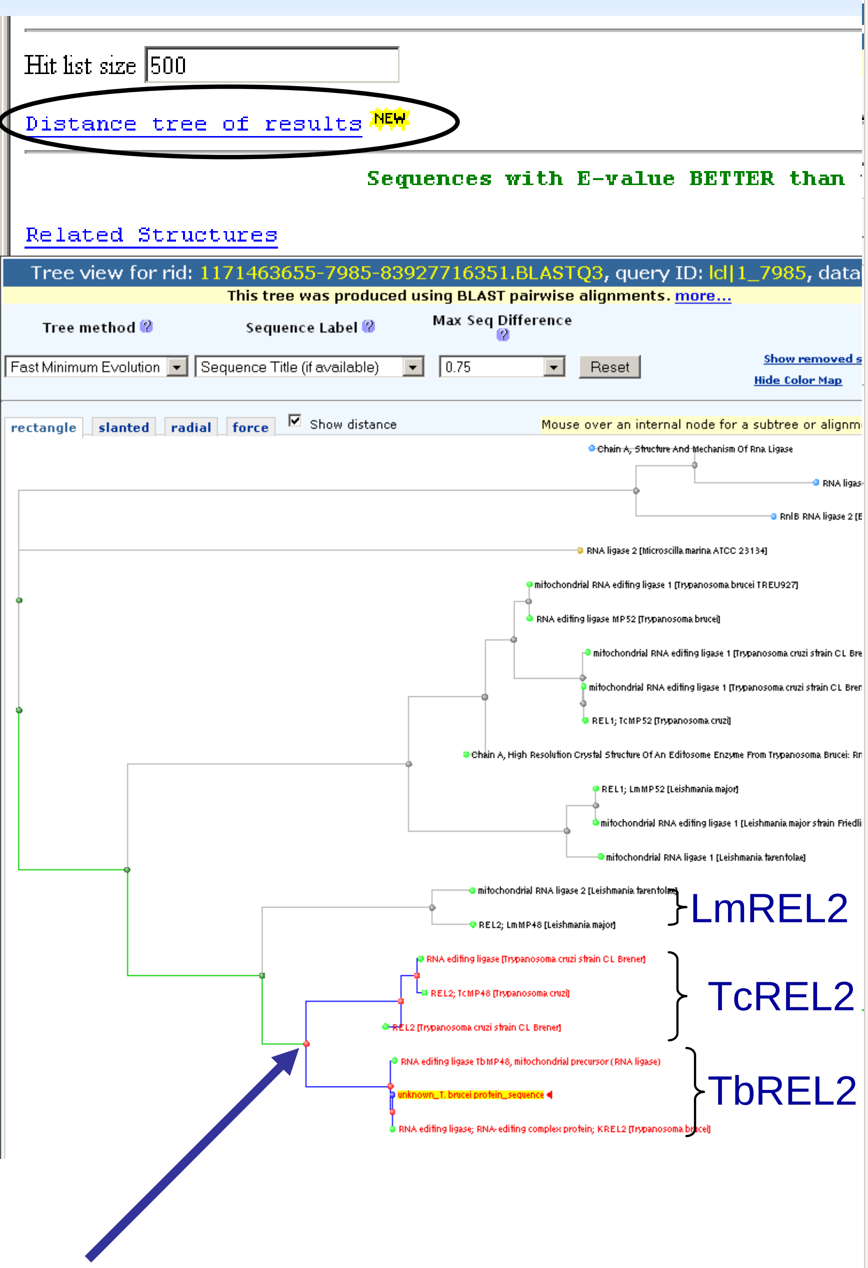

[Gly]

- V364A (nonpolar, aliphatic, hydrophobic to tiny, nonpolar, aliphatic) conservative


[Val]


[Ala]

Example 1: check distance tree and alignments from NCBI BLAST output



Hit list size 500

Distance tree of results NEW

Sequences with E-value BETTER than

Related Structures

Tree view for rid: 1171463655-7985-83927716351.BLASTQ3, query ID: lcl|1_7985, data

This tree was produced using BLAST pairwise alignments. more...

Tree method: Fast Minimum Evolution

Sequence Label: Sequence Title (if available)

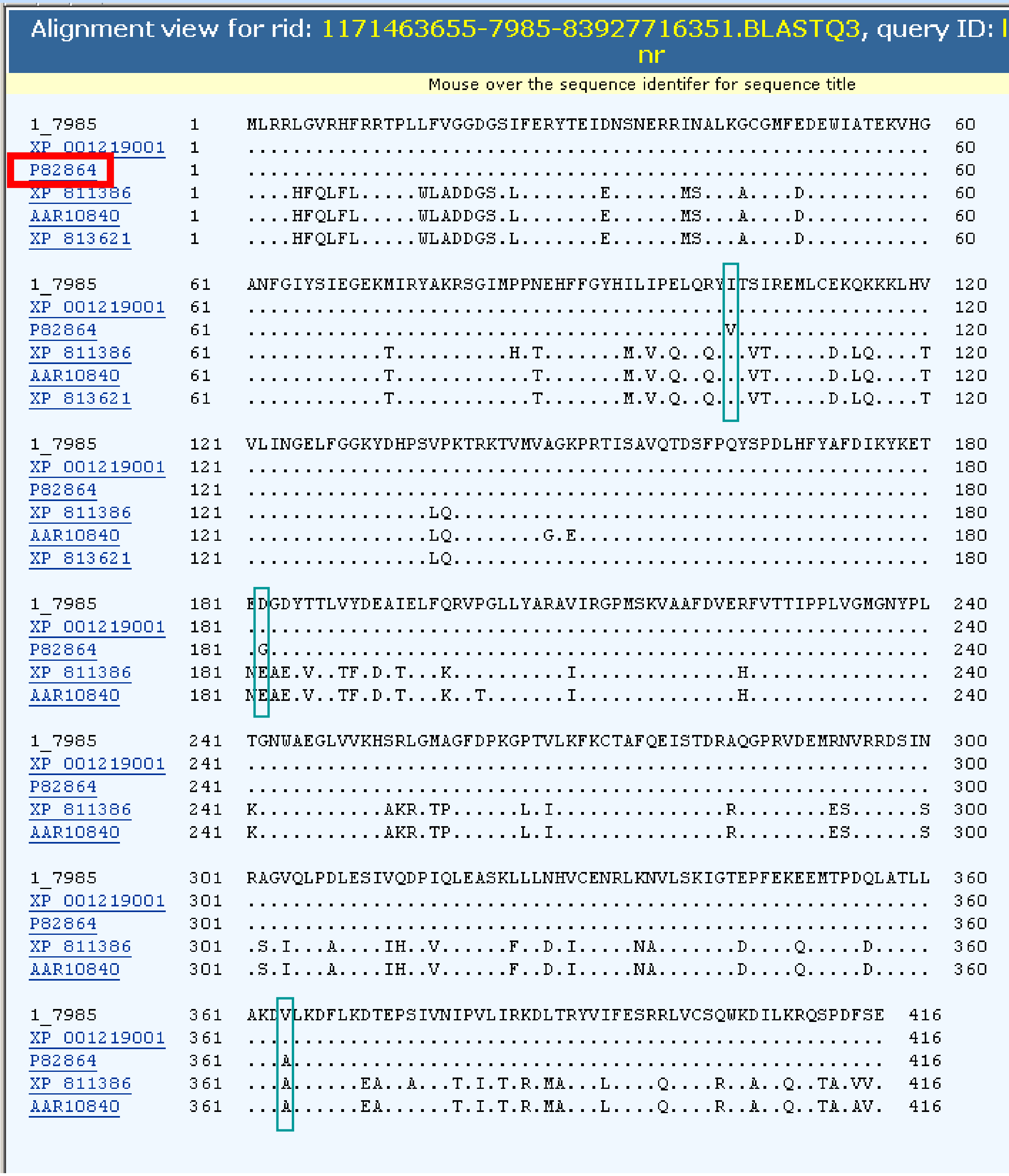
Max Seq Difference: 0.75

☑ Show distance

Mouse over an internal node for a subtree or alignm

Reset
Show removed s
Hide Color Map
rectangle
slanted
radial
force
Chain A, Structure And Mechanism Of Rna Ligase
RNA ligas
RnlB RNA ligase 2 [E
RNA ligase 2 [Microscilla marina ATCC 23134]
mitochondrial RNA editing ligase 1 [Trypanosoma brucei TREU927]
RNA editing ligase MP52 [Trypanosoma brucei]
mitochondrial RNA editing ligase 1 [Trypanosoma cruzi strain CL Bre
mitochondrial RNA editing ligase 1 [Trypanosoma cruzi strain CL Bren
REL1; TcMP52 [Trypanosoma cruzi]
Chain A, High Resolution Crystal Structure Of An Editosome Enzyme From Trypanosoma Brucei: Rn
REL1; LmMP52 [Leishmania major]
mitochondrial RNA editing ligase 1 [Leishmania major strain Friedli
mitochondrial RNA ligase 1 [Leishmania tarentolae]
mitochondrial RNA ligase 2 [Leishmania tarentolae]
REL2; LmMP48 [Leishmania major]
RNA editing ligase [Trypanosoma cruzi strain CL Brener]
REL2; TcMP48 [Trypanosoma cruzi]
REL2 [Trypanosoma cruzi strain CL Brener]
RNA editing ligase TbMP48, mitochondrial precursor (RNA ligase)
unknown_T. brucei protein_sequence
RNA editing ligase; RNA-editing complex protein; KREL2 [Trypanosoma brucei]
LmREL2
TcREL2
TbREL2

Click here at the branch point



Alignment view for rid: 1171463655-7985-83927716351.BLASTQ3, query ID: l nr

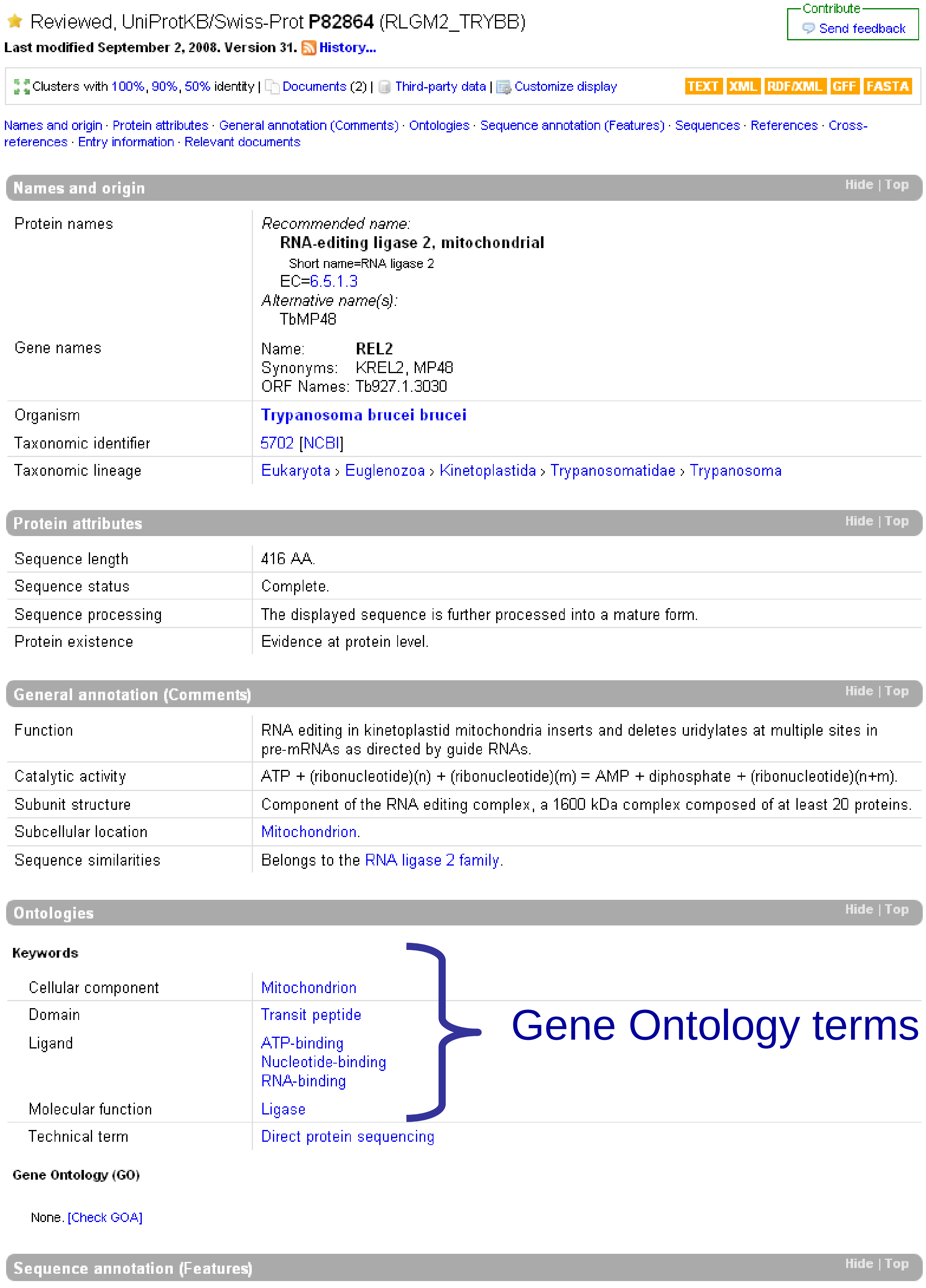
Mouse over the sequence identifer for sequence title

```
1_7985          1    MLRRLGVRHFRRTPLLFVGGDGSIFERYTEIDNSNERRINALKGCGMFEDEWIATEKVHG  60
XP_001219001    1    ............................................................  60
P82864          1    ............................................................  60
XP_811386       1    ....HFQLFL.....WLADDGS.L.......E......MS...A....D...........  60
AAR10840        1    ....HFQLFL.....WLADDGS.L.......E......MS...A....D...........  60
XP_813621       1    ....HFQLFL.....WLADDGS.L.......E......MS...A....D...........  60

1_7985          61   ANFGIYSIEGEKMIRYAKRSGIMPPNEHFFGYHILIPELQRYITSIREMLCEKQKKKLHV  120
XP_001219001    61   ............................................................  120
P82864          61   ..........................................V.................  120
XP_811386       61   ............T..........H.T.......M.V.Q..Q...VT.....D.LQ....T  120
AAR10840        61   ............T............T.......M.V.Q..Q...VT.....D.LQ....T  120
XP_813621       61   ............T............T.......M.V.Q..Q...VT.....D.LQ....T  120

1_7985          121  VLINGELFGGKYDHPSVPKTRKTVMVAGKPRTISAVQTDSFPQYSPDLHFYAFDIKYKET  180
XP_001219001    121  ............................................................  180
P82864          121  ............................................................  180
XP_811386       121  ................LQ..........................................  180
AAR10840        121  ................LQ........G.E...............................  180
XP_813621       121  ................LQ..........................................  180

1_7985          181  EDGDYTTLVYDEAIELFQRVPGLLYARAVIRGPMSKVAAFDVERFVTTIPPLVGMGNYPL  240
XP_001219001    181  ............................................................  240
P82864          181  .G..........................................................  240
XP_811386       181  NEAE.V..TF.D.T...K..........I..............H................  240
AAR10840        181  NEAE.V..TF.D.T...K..T.......I..............H................  240

1_7985          241  TGNWAEGLVVKHSRLGMAGFDPKGPTVLKFKCTAFQEISTDRAQGPRVDEMRNVRRDSIN  300
XP_001219001    241  ............................................................  300
P82864          241  ............................................................  300
XP_811386       241  K...........AKR.TP......L.I...............R........ES......S  300
AAR10840        241  K...........AKR.TP......L.I...............R........ES......S  300

1_7985          301  RAGVQLPDLESIVQDPIQLEASKLLLNHVCENRLKNVLSKIGTEPFEKEEMTPDQLATLL  360
XP_001219001    301  ............................................................  360
P82864          301  ............................................................  360
XP_811386       301  .S.I...A....IH..V......F..D.I.....NA.......D....Q.....D.....  360
AAR10840        301  .S.I...A....IH..V......F..D.I.....NA.......D....Q.....D.....  360

1_7985          361  AKDVLKDFLKDTEPSIVNIPVLIRKDLTRYVIFESRRLVCSQWKDILKRQSPDFSE  416
XP_001219001    361  ........................................................  416
P82864          361  ...A....................................................  416
XP_811386       361  ...A......EA..A...T.I.T.R.MA...L....Q....R..A..Q..TA.VV.  416
AAR10840        361  ...A......EA......T.I.T.R.MA...L....Q....R..A..Q..TA.AV.  416
```

Swiss-Prot

The protein sequence is 99% identical to the sequence of this Swiss-Prot entry, P82864. Protein name is “RNA-editing ligase 2, mitochondrial.”

Gene name is ‘REL2.’



Reviewed, UniProtKB/Swiss-Prot **P82864** (RLGM2_TRYBB)

Last modified September 2, 2008. Version 31. **History...**

Contribute
Send feedback

Clusters with 100%, 90%, 50% identity | Documents (2) | Third-party data | Customize display

TEXT XML RDF/XML GFF FASTA

Names and origin · Protein attributes · General annotation (Comments) · Ontologies · Sequence annotation (Features) · Sequences · References · Cross-references · Entry information · Relevant documents

Names and origin Hide | Top

Protein names	*Recommended name:* **RNA-editing ligase 2, mitochondrial** Short name=RNA ligase 2 EC=6.5.1.3 *Alternative name(s):* TbMP48
Gene names	Name: **REL2** Synonyms: KREL2, MP48 ORF Names: Tb927.1.3030
Organism	**Trypanosoma brucei brucei**
Taxonomic identifier	5702 [NCBI]
Taxonomic lineage	Eukaryota › Euglenozoa › Kinetoplastida › Trypanosomatidae › Trypanosoma

Protein attributes Hide | Top

Sequence length	416 AA.
Sequence status	Complete.
Sequence processing	The displayed sequence is further processed into a mature form.
Protein existence	Evidence at protein level.

General annotation (Comments) Hide | Top

Function	RNA editing in kinetoplastid mitochondria inserts and deletes uridylates at multiple sites in pre-mRNAs as directed by guide RNAs.
Catalytic activity	ATP + (ribonucleotide)(n) + (ribonucleotide)(m) = AMP + diphosphate + (ribonucleotide)(n+m).
Subunit structure	Component of the RNA editing complex, a 1600 kDa complex composed of at least 20 proteins.
Subcellular location	Mitochondrion.
Sequence similarities	Belongs to the RNA ligase 2 family.

Ontologies Hide | Top

Keywords

Cellular component	Mitochondrion
Domain	Transit peptide
Ligand	ATP-binding Nucleotide-binding RNA-binding
Molecular function	Ligase
Technical term	Direct protein sequencing

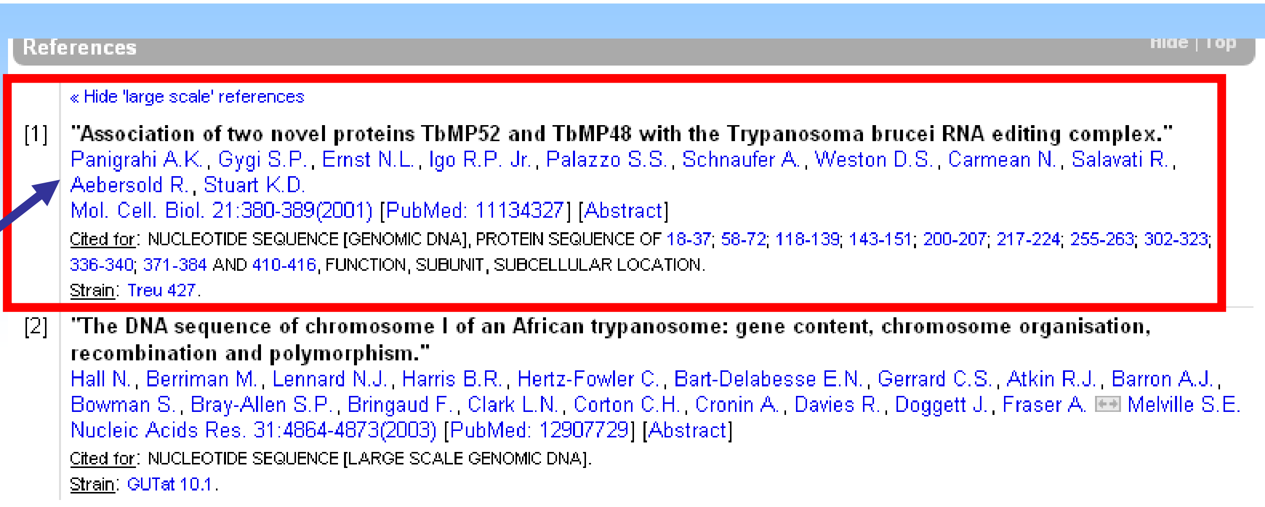
Gene Ontology (GO)

None. [Check GOA]

Sequence annotation (Features) Hide | Top

Gene Ontology terms

Second Swiss-Prot Page

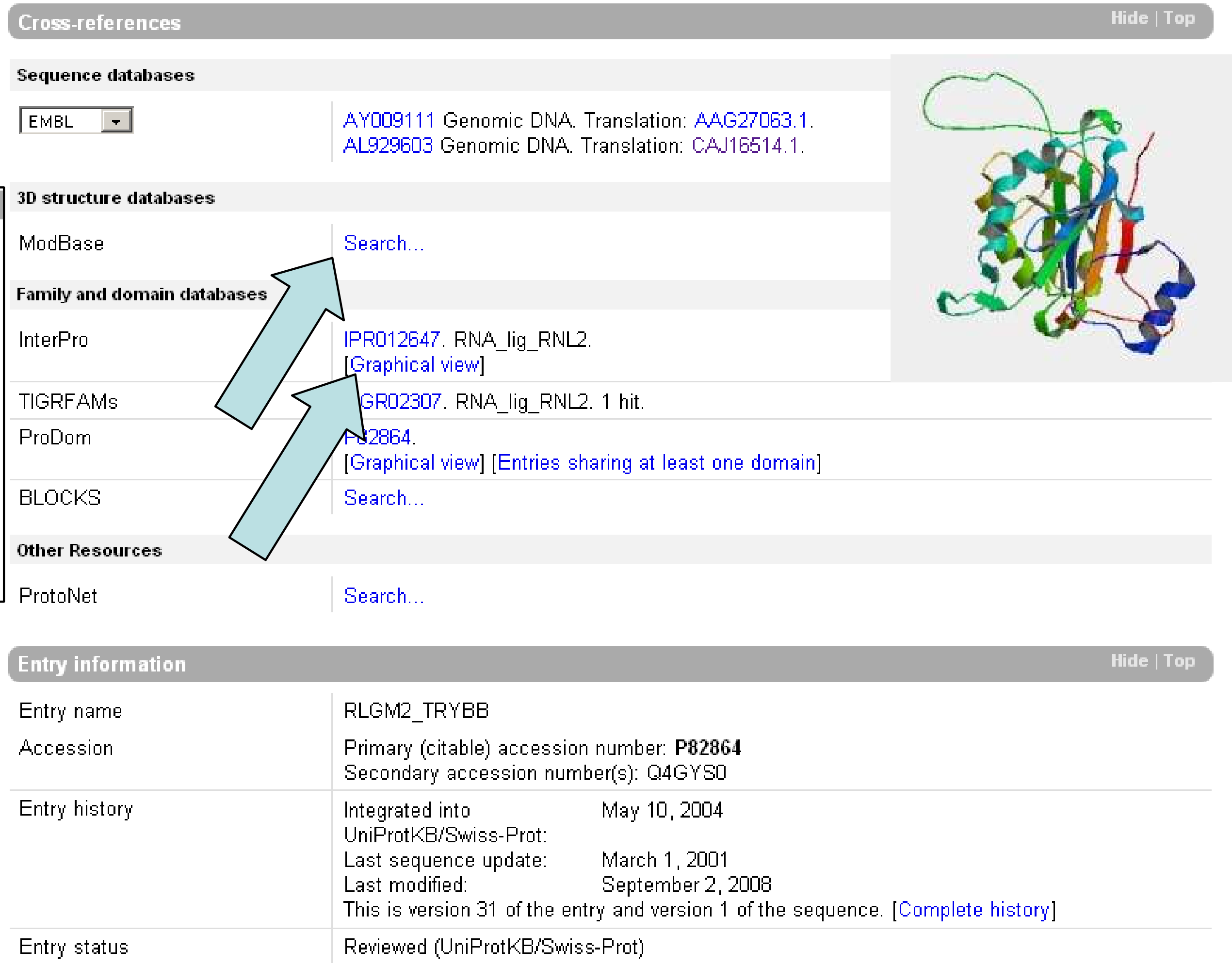


« Hide 'large scale' references

[1] **"Association of two novel proteins TbMP52 and TbMP48 with the Trypanosoma brucei RNA editing complex."**
Panigrahi A.K., Gygi S.P., Ernst N.L., Igo R.P. Jr., Palazzo S.S., Schnaufer A., Weston D.S., Carmean N., Salavati R., Aebersold R., Stuart K.D.
Mol. Cell. Biol. 21:380-389(2001) [PubMed: 11134327] [Abstract]
Cited for: NUCLEOTIDE SEQUENCE [GENOMIC DNA], PROTEIN SEQUENCE OF 18-37; 58-72; 118-139; 143-151; 200-207; 217-224; 255-263; 302-323; 336-340; 371-384 AND 410-416, FUNCTION, SUBUNIT, SUBCELLULAR LOCATION.
Strain: Treu 427.

[2] **"The DNA sequence of chromosome I of an African trypanosome: gene content, chromosome organisation, recombination and polymorphism."**
Hall N., Berriman M., Lennard N.J., Harris B.R., Hertz-Fowler C., Bart-Delabesse E.N., Gerrard C.S., Atkin R.J., Barron A.J., Bowman S., Bray-Allen S.P., Bringaud F., Clark L.N., Corton C.H., Cronin A., Davies R., Doggett J., Fraser A. ↔ Melville S.E.
Nucleic Acids Res. 31:4864-4873(2003) [PubMed: 12907729] [Abstract]
Cited for: NUCLEOTIDE SEQUENCE [LARGE SCALE GENOMIC DNA].
Strain: GUTat 10.1.

Click on the hyperlink to look at this publication.



Sequence databases

EMBL	AY009111 Genomic DNA. Translation: AAG27063.1. AL929603 Genomic DNA. Translation: CAJ16514.1.

3D structure databases

ModBase	Search...

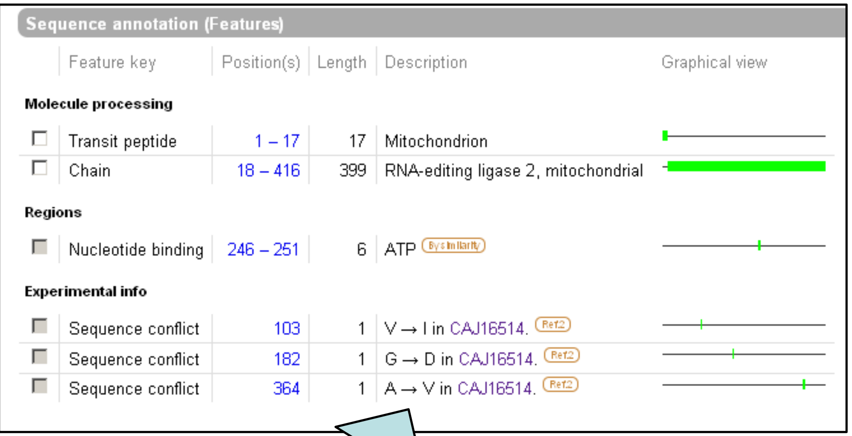
Family and domain databases

InterPro	IPR012647. RNA_lig_RNL2. [Graphical view]
TIGRFAMs	GR02307. RNA_lig_RNL2. 1 hit.
ProDom	P2864. [Graphical view] [Entries sharing at least one domain]
BLOCKS	Search...

Other Resources

ProtoNet	Search...

Entry name	RLGM2_TRYBB
Accession	Primary (citable) accession number: **P82864** Secondary accession number(s): Q4GYS0
Entry history	Integrated into UniProtKB/Swiss-Prot: May 10, 2004 Last sequence update: March 1, 2001 Last modified: September 2, 2008 This is version 31 of the entry and version 1 of the sequence. [Complete history]
Entry status	Reviewed (UniProtKB/Swiss-Prot)

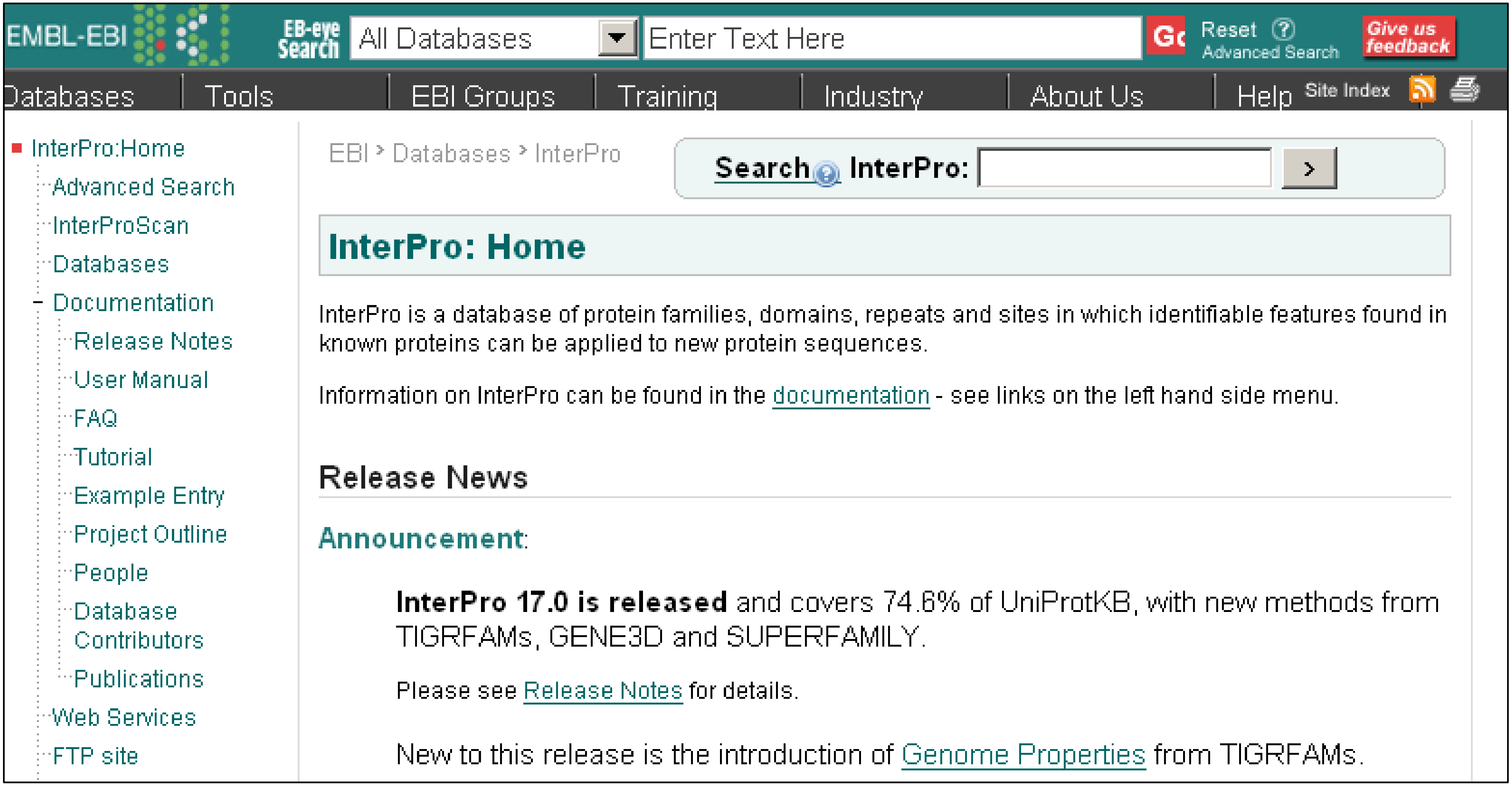


	Feature key	Position(s)	Length	Description	Graphical view
Molecule processing					
□	Transit peptide	1 – 17	17	Mitochondrion	
□	Chain	18 – 416	399	RNA-editing ligase 2, mitochondrial	
Regions					
□	Nucleotide binding	246 – 251	6	ATP (By similarity)	
Experimental info					
□	Sequence conflict	103	1	V → I in CAJ16514. (Ref.2)	
□	Sequence conflict	182	1	G → D in CAJ16514. (Ref.2)	
□	Sequence conflict	364	1	A → V in CAJ16514. (Ref.2)	

The three SNPs we noted are noted here in the Swiss-Prot record.

Interpro

http://www.ebi.ac.uk/interpro/



InterPro: Home

InterPro is a database of protein families, domains, repeats and sites in which identifiable features found in known proteins can be applied to new protein sequences.

Information on InterPro can be found in the documentation - see links on the left hand side menu.

Release News

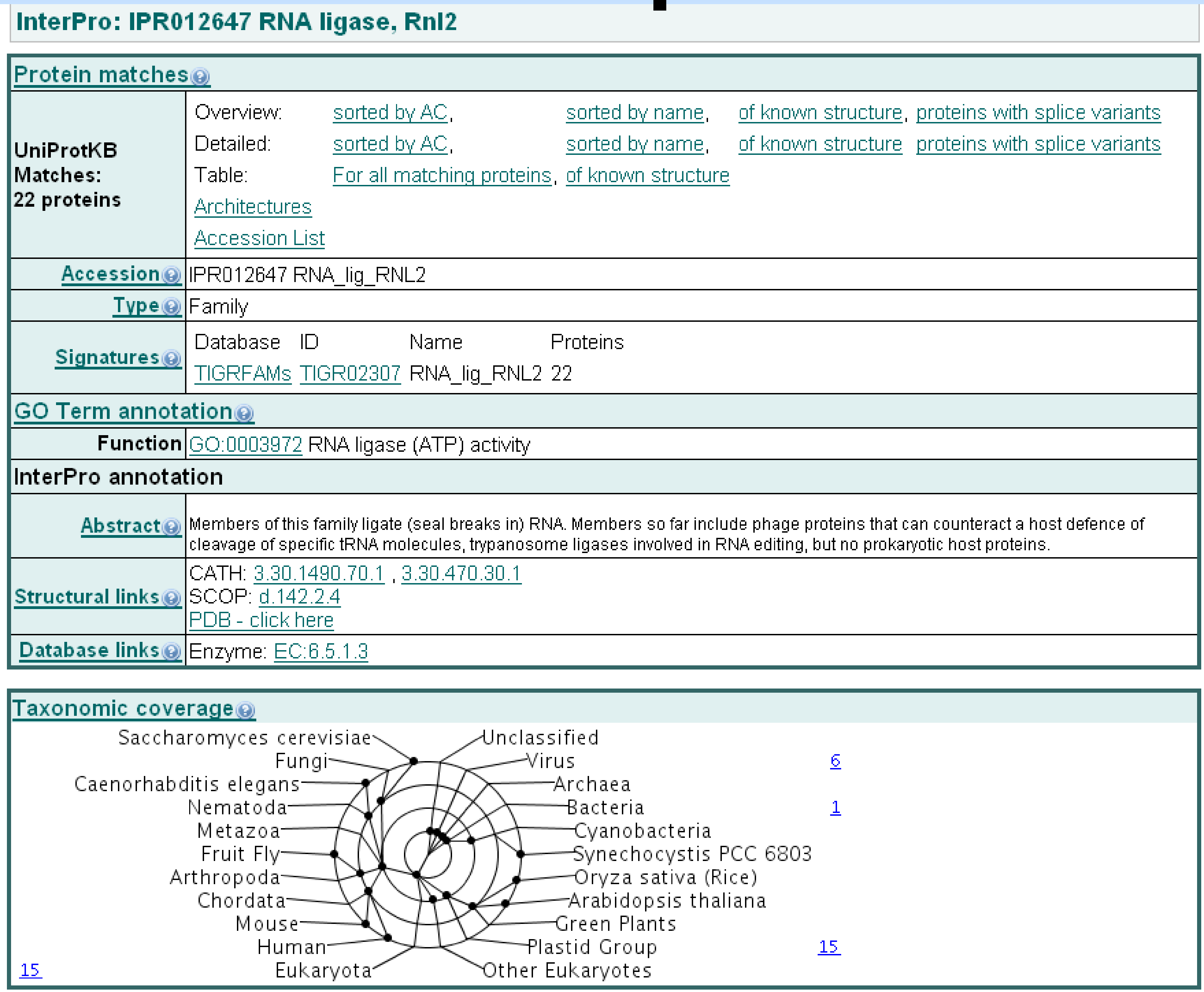
Announcement:

InterPro 17.0 is released and covers 74.6% of UniProtKB, with new methods from TIGRFAMs, GENE3D and SUPERFAMILY.

Please see Release Notes for details.

New to this release is the introduction of Genome Properties from TIGRFAMs.

Interpro result



InterPro: IPR012647 RNA ligase, Rnl2

Protein matches

UniProtKB Matches: 22 proteins	Overview: sorted by AC, sorted by name, of known structure, proteins with splice variants Detailed: sorted by AC, sorted by name, of known structure proteins with splice variants Table: For all matching proteins, of known structure Architectures Accession List
Accession	IPR012647 RNA_lig_RNL2
Type	Family
Signatures	Database ID Name Proteins TIGRFAMs TIGR02307 RNA_lig_RNL2 22

GO Term annotation

Function	GO:0003972 RNA ligase (ATP) activity

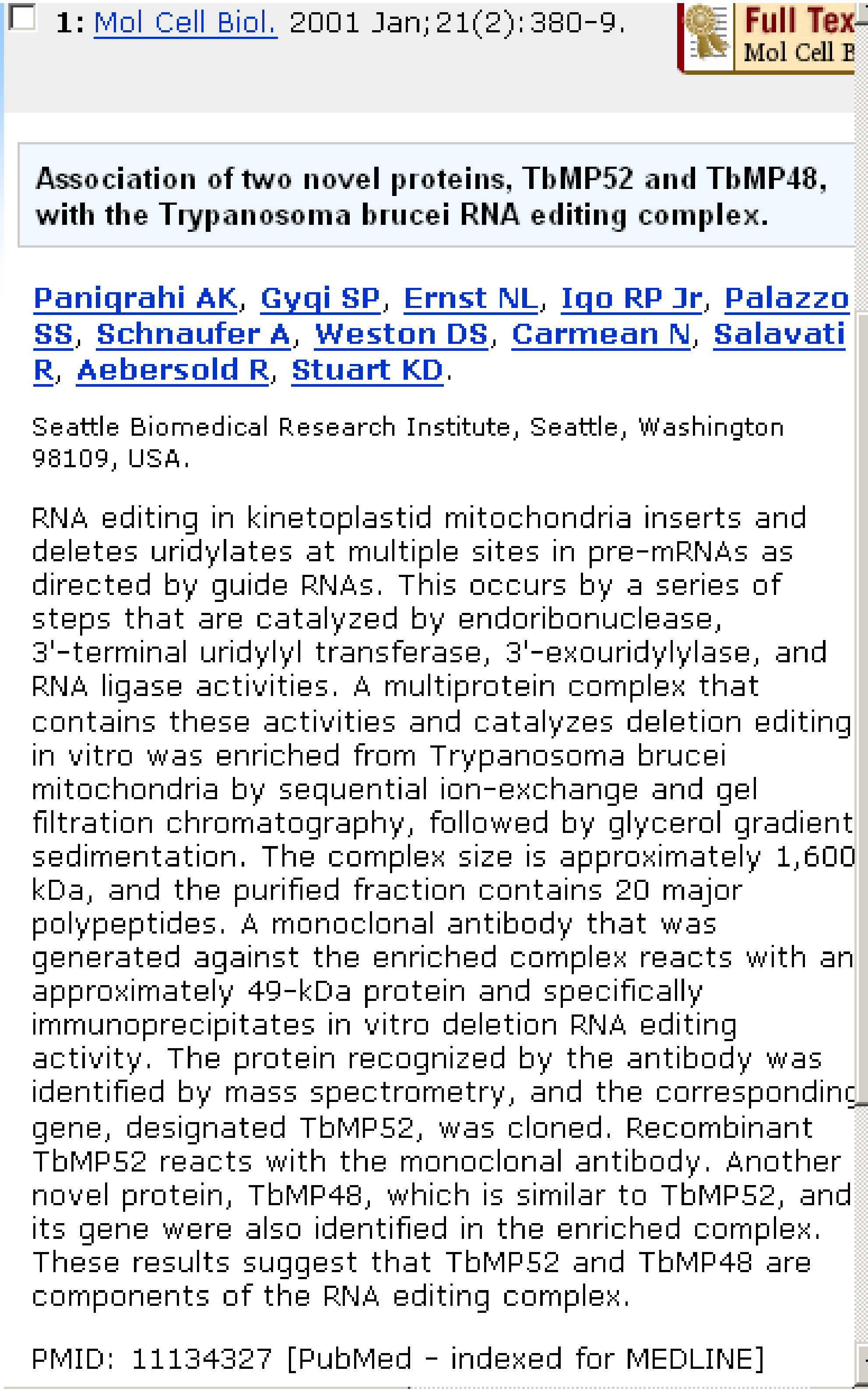
InterPro annotation

Abstract	Members of this family ligate (seal breaks in) RNA. Members so far include phage proteins that can counteract a host defence of cleavage of specific tRNA molecules, trypanosome ligases involved in RNA editing, but no prokaryotic host proteins.
Structural links	CATH: 3.30.1490.70.1 , 3.30.470.30.1 SCOP: d.142.2.4 PDB - click here
Database links	Enzyme: EC:6.5.1.3

Taxonomic coverage

Pubmed

- Read the abstract.
- If promising, read the paper to be sure protein is characterized.
- If characterized, it is good evidence for naming our protein sequence.



□ 1: Mol Cell Biol. 2001 Jan;21(2):380-9.

Association of two novel proteins, TbMP52 and TbMP48, with the Trypanosoma brucei RNA editing complex.

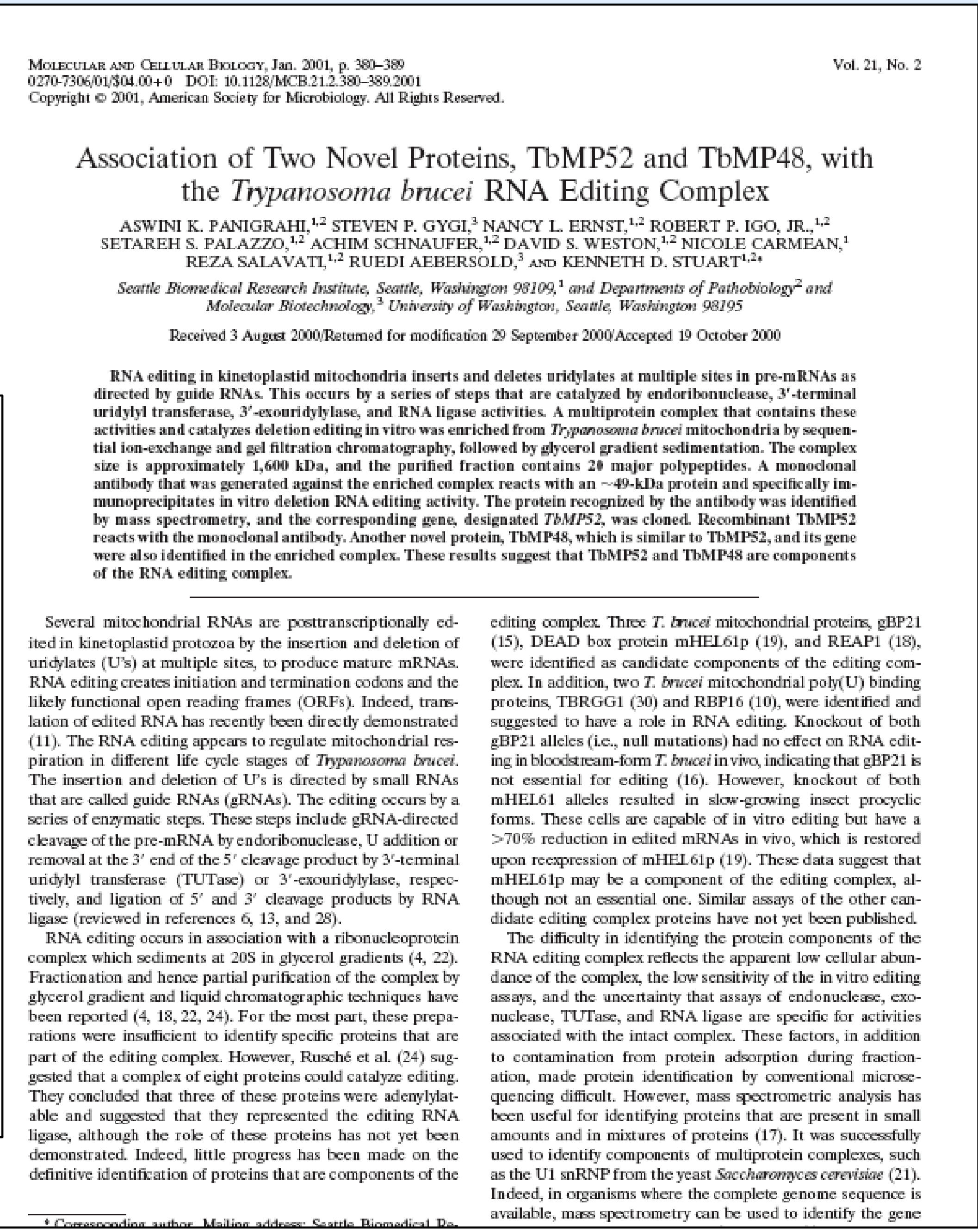
Panigrahi AK, Gygi SP, Ernst NL, Igo RP Jr, Palazzo SS, Schnaufer A, Weston DS, Carmean N, Salavati R, Aebersold R, Stuart KD.

Seattle Biomedical Research Institute, Seattle, Washington 98109, USA.

RNA editing in kinetoplastid mitochondria inserts and deletes uridylates at multiple sites in pre-mRNAs as directed by guide RNAs. This occurs by a series of steps that are catalyzed by endoribonuclease, 3'-terminal uridylyl transferase, 3'-exouridylylase, and RNA ligase activities. A multiprotein complex that contains these activities and catalyzes deletion editing in vitro was enriched from Trypanosoma brucei mitochondria by sequential ion-exchange and gel filtration chromatography, followed by glycerol gradient sedimentation. The complex size is approximately 1,600 kDa, and the purified fraction contains 20 major polypeptides. A monoclonal antibody that was generated against the enriched complex reacts with an approximately 49-kDa protein and specifically immunoprecipitates in vitro deletion RNA editing activity. The protein recognized by the antibody was identified by mass spectrometry, and the corresponding gene, designated TbMP52, was cloned. Recombinant TbMP52 reacts with the monoclonal antibody. Another novel protein, TbMP48, which is similar to TbMP52, and its gene were also identified in the enriched complex. These results suggest that TbMP52 and TbMP48 are components of the RNA editing complex.

PMID: 11134327 [PubMed - indexed for MEDLINE]

The paper



MOLECULAR AND CELLULAR BIOLOGY, Jan. 2001, p. 380–389 Vol. 21, No. 2
0270-7306/01/$04.00+0 DOI: 10.1128/MCB.21.2.380–389.2001

Association of Two Novel Proteins, TbMP52 and TbMP48, with the *Trypanosoma brucei* RNA Editing Complex

ASWINI K. PANIGRAHI,[1,2] STEVEN P. GYGI,[3] NANCY L. ERNST,[1,2] ROBERT P. IGO, JR.,[1,2]
SETAREH S. PALAZZO,[1,2] ACHIM SCHNAUFER,[1,2] DAVID S. WESTON,[1,2] NICOLE CARMEAN,[1]
REZA SALAVATI,[1,2] RUEDI AEBERSOLD,[3] AND KENNETH D. STUART[1,2*]

Seattle Biomedical Research Institute, Seattle, Washington 98109,[1] and Departments of Pathobiology[2] and Molecular Biotechnology,[3] University of Washington, Seattle, Washington 98195

Received 3 August 2000/Returned for modification 29 September 2000/Accepted 19 October 2000

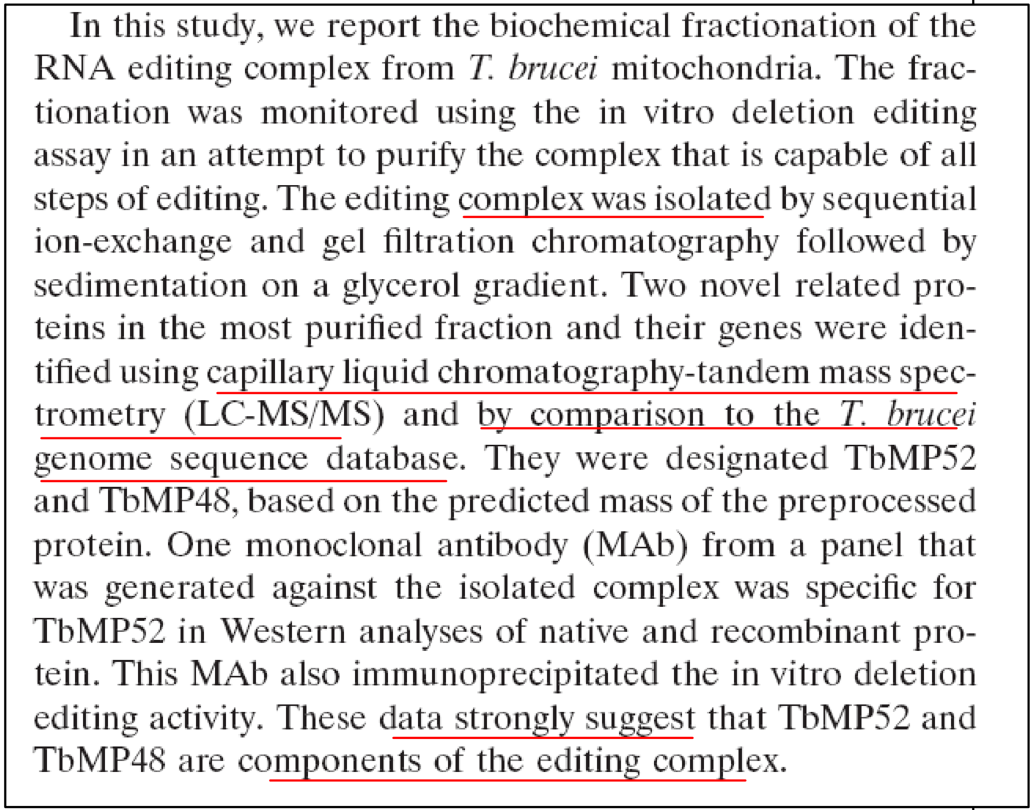
RNA editing in kinetoplastid mitochondria inserts and deletes uridylates at multiple sites in pre-mRNAs as directed by guide RNAs. This occurs by a series of steps that are catalyzed by endoribonuclease, 3′-terminal uridylyl transferase, 3′-exouridylylase, and RNA ligase activities. A multiprotein complex that contains these activities and catalyzes deletion editing in vitro was enriched from *Trypanosoma brucei* mitochondria by sequential ion-exchange and gel filtration chromatography, followed by glycerol gradient sedimentation. The complex size is approximately 1,600 kDa, and the purified fraction contains 20 major polypeptides. A monoclonal antibody that was generated against the enriched complex reacts with an ~49-kDa protein and specifically immunoprecipitates in vitro deletion RNA editing activity. The protein recognized by the antibody was identified by mass spectrometry, and the corresponding gene, designated *TbMP52*, was cloned. Recombinant TbMP52 reacts with the monoclonal antibody. Another novel protein, TbMP48, which is similar to TbMP52, and its gene were also identified in the enriched complex. These results suggest that TbMP52 and TbMP48 are components of the RNA editing complex.

Several mitochondrial RNAs are posttranscriptionally edited in kinetoplastid protozoa by the insertion and deletion of uridylates (U's) at multiple sites, to produce mature mRNAs. RNA editing creates initiation and termination codons and the likely functional open reading frames (ORFs). Indeed, translation of edited RNA has recently been directly demonstrated (11). The RNA editing appears to regulate mitochondrial respiration in different life cycle stages of *Trypanosoma brucei*. The insertion and deletion of U's is directed by small RNAs that are called guide RNAs (gRNAs). The editing occurs by a series of enzymatic steps. These steps include gRNA-directed cleavage of the pre-mRNA by endoribonuclease, U addition or removal at the 3′ end of the 5′ cleavage product by 3′-terminal uridylyl transferase (TUTase) or 3′-exouridylylase, respectively, and ligation of 5′ and 3′ cleavage products by RNA ligase (reviewed in references 6, 13, and 28).

RNA editing occurs in association with a ribonucleoprotein complex which sediments at 20S in glycerol gradients (4, 22). Fractionation and hence partial purification of the complex by glycerol gradient and liquid chromatographic techniques have been reported (4, 18, 22, 24). For the most part, these preparations were insufficient to identify specific proteins that are part of the editing complex. However, Rusché et al. (24) suggested that a complex of eight proteins could catalyze editing. They concluded that three of these proteins were adenylylatable and suggested that they represented the editing RNA ligase, although the role of these proteins has not yet been demonstrated. Indeed, little progress has been made on the definitive identification of proteins that are components of the editing complex. Three *T. brucei* mitochondrial proteins, gBP21 (15), DEAD box protein mHEL61p (19), and REAP1 (18), were identified as candidate components of the editing complex. In addition, two *T. brucei* mitochondrial poly(U) binding proteins, TBRGG1 (30) and RBP16 (10), were identified and suggested to have a role in RNA editing. Knockout of both gBP21 alleles (i.e., null mutations) had no effect on RNA editing in bloodstream-form *T. brucei* in vivo, indicating that gBP21 is not essential for editing (16). However, knockout of both mHEL61 alleles resulted in slow-growing insect procyclic forms. These cells are capable of in vitro editing but have a >70% reduction in edited mRNAs in vivo, which is restored upon reexpression of mHEL61p (19). These data suggest that mHEL61p may be a component of the editing complex, although not an essential one. Similar assays of the other candidate editing complex proteins have not yet been published.

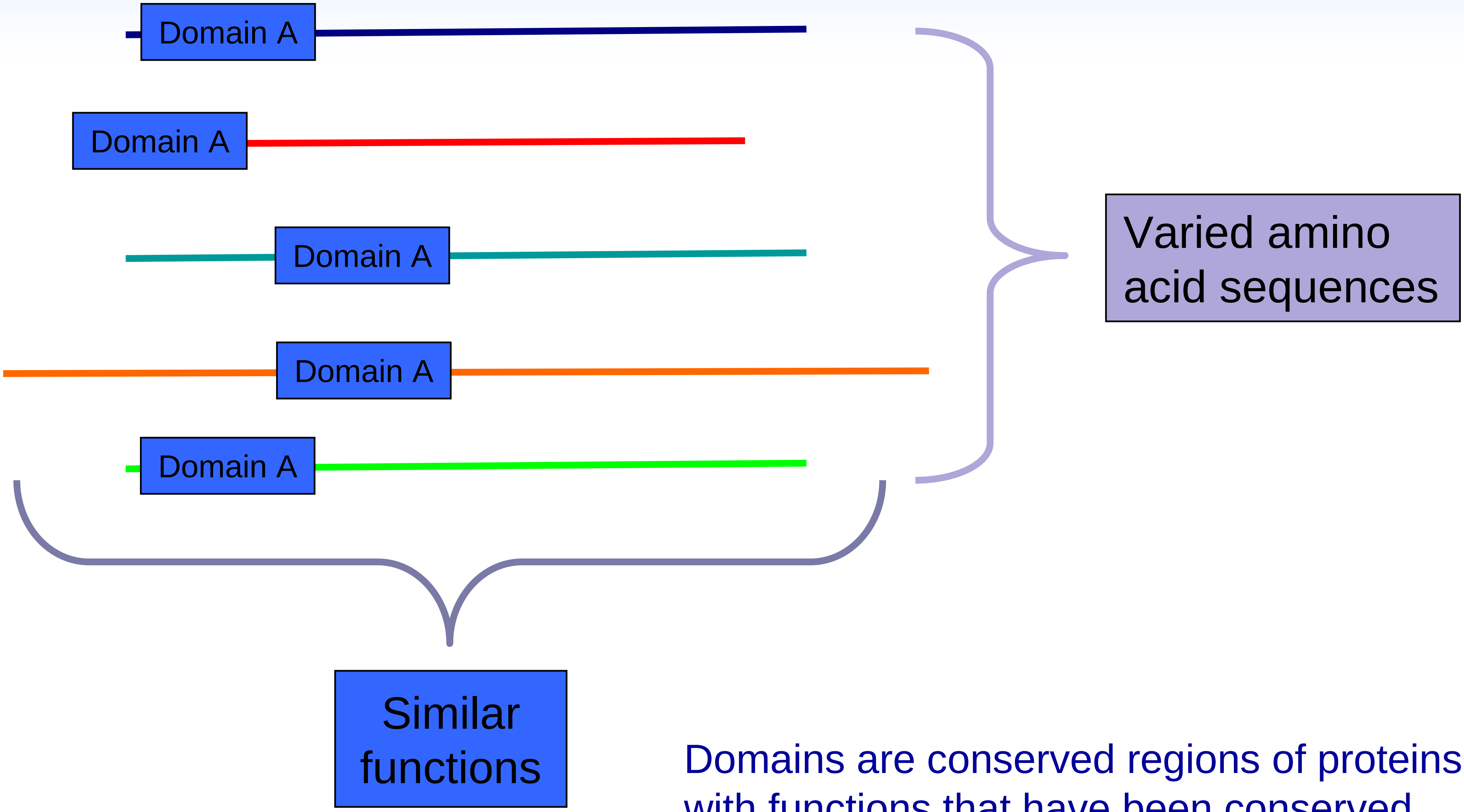
The difficulty in identifying the protein components of the RNA editing complex reflects the apparent low cellular abundance of the complex, the low sensitivity of the in vitro editing assays, and the uncertainty that assays of endonuclease, exonuclease, TUTase, and RNA ligase are specific for activities associated with the intact complex. These factors, in addition to contamination from protein adsorption during fractionation, made protein identification by conventional microsequencing difficult. However, mass spectrometric analysis has been useful for identifying proteins that are present in small amounts and in mixtures of proteins (17). It was successfully used to identify components of multiprotein complexes, such as the U1 snRNP from the yeast *Saccharomyces cerevisiae* (21). Indeed, in organisms where the complete genome sequence is available, mass spectrometry can be used to identify the gene

* Corresponding author. Mailing address: Seattle Biomedical Re-



In this study, we report the biochemical fractionation of the RNA editing complex from *T. brucei* mitochondria. The fractionation was monitored using the in vitro deletion editing assay in an attempt to purify the complex that is capable of all steps of editing. The editing complex was isolated by sequential ion-exchange and gel filtration chromatography followed by sedimentation on a glycerol gradient. Two novel related proteins in the most purified fraction and their genes were identified using capillary liquid chromatography-tandem mass spectrometry (LC-MS/MS) and by comparison to the *T. brucei* genome sequence database. They were designated TbMP52 and TbMP48, based on the predicted mass of the preprocessed protein. One monoclonal antibody (MAb) from a panel that was generated against the isolated complex was specific for TbMP52 in Western analyses of native and recombinant protein. This MAb also immunoprecipitated the in vitro deletion editing activity. These data strongly suggest that TbMP52 and TbMP48 are components of the editing complex.

HMMs



Domains are conserved regions of proteins with functions that have been conserved throughout evolution.

Pfam

http://pfam.janiela.org/

Pfam is a large collection of protein families. The families are built around domain composition. Domains are computed from multiple sequence alignments that are used to generate hidden Markov models.

For each family in Pfam you can:

- Look at multiple alignments
- View protein domain architectures
- Examine species distribution
- Follow links to other databases
- View known protein structures

TIGRfams

http//www.tigr.org/TIGRFAMs/index.shtml

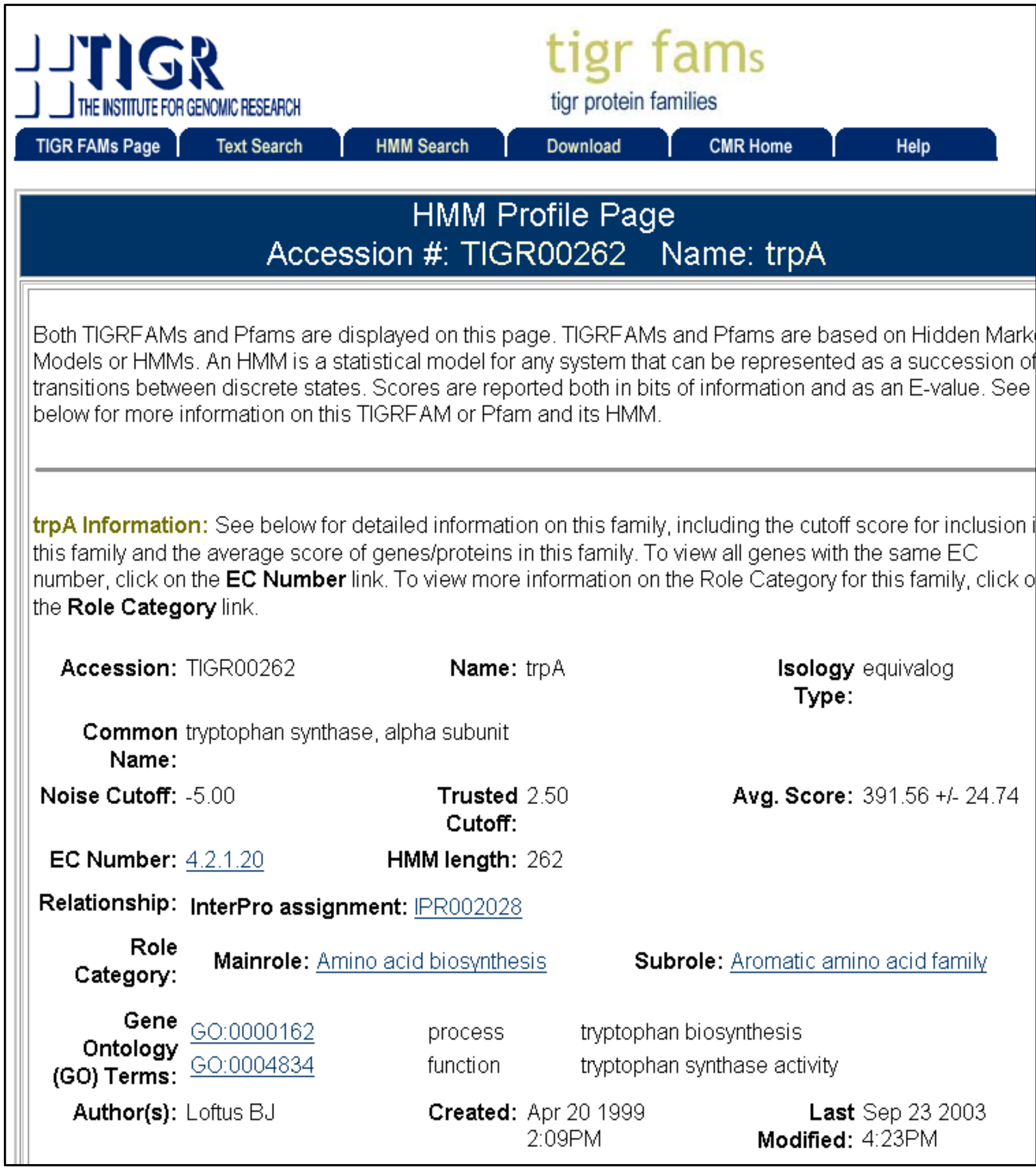
TIGRFAMs: a collection of protein families featuring curated multiple sequence alignments, Hidden Markov Models (HMMs) and associated information designed to support the automated functional identification of proteins by sequence homology. Use the TIGRfam page to see

- the curated seed alignment for each TIGRFAM
- the full alignment of all family members
- the cutoff scores for inclusion in each of the TIGRfams.

Also use this page to search through the TIGRfams and HMMs

•for text (TIGRfams Text Search) or

•for specific sequences (TIGRfams Sequence Search).



HMM Profile Page
Accession #: TIGR00262 Name: trpA

Both TIGRFAMs and Pfams are displayed on this page. TIGRFAMs and Pfams are based on Hidden Mark Models or HMMs. An HMM is a statistical model for any system that can be represented as a succession of transitions between discrete states. Scores are reported both in bits of information and as an E-value. See below for more information on this TIGRFAM or Pfam and its HMM.

trpA Information: See below for detailed information on this family, including the cutoff score for inclusion i this family and the average score of genes/proteins in this family. To view all genes with the same EC number, click on the **EC Number** link. To view more information on the Role Category for this family, click o the **Role Category** link.

Accession: TIGR00262 **Name:** trpA **Isology Type:** equivalog

Common Name: tryptophan synthase, alpha subunit

Noise Cutoff: -5.00 **Trusted Cutoff:** 2.50 **Avg. Score:** 391.56 +/- 24.74

EC Number: 4.2.1.20 **HMM length:** 262

Relationship: **InterPro assignment:** IPR002028

Role Category: **Mainrole:** Amino acid biosynthesis **Subrole:** Aromatic amino acid family

Gene Ontology (GO) Terms:		
GO:0000162	process	tryptophan biosynthesis
GO:0004834	function	tryptophan synthase activity

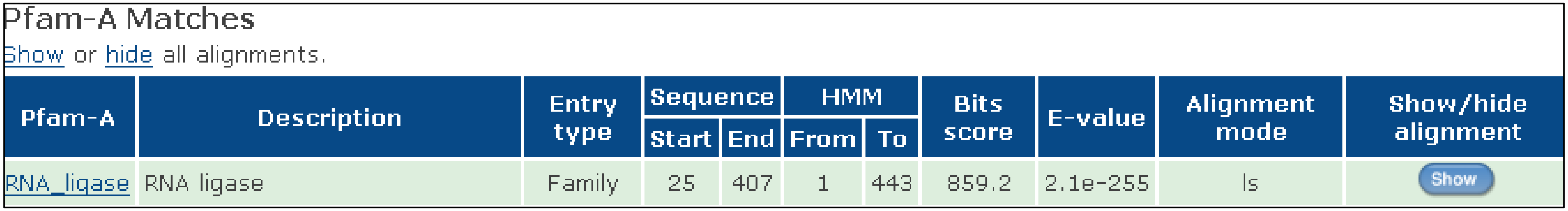
Author(s): Loftus BJ **Created:** Apr 20 1999 2:09PM **Last Modified:** Sep 23 2003 4:23PM

Domain results

Pfam search:
pfam.janiela.org

Total score: 859.2
E-value: 2.1 e-255

This is a very positive hit to the RNA ligase domain.

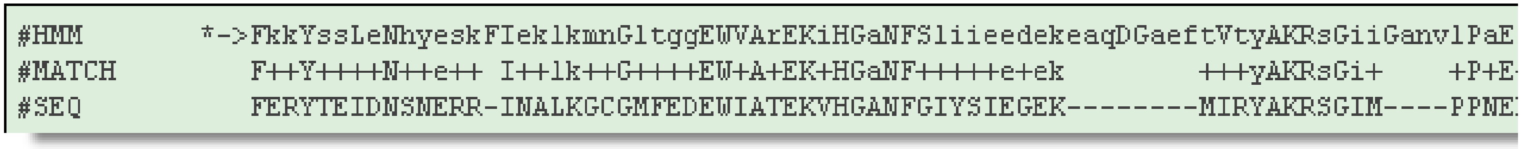


Pfam-A Matches

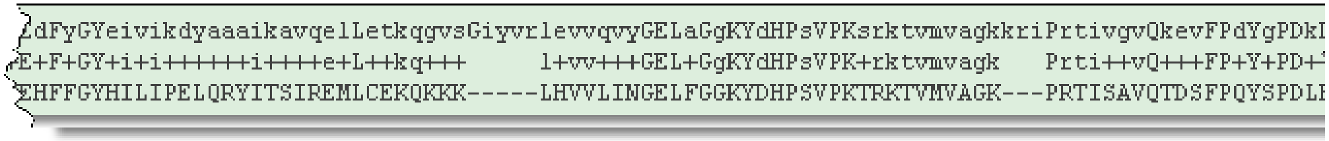
Show or hide all alignments.

Pfam-A	Description	Entry type	Sequence		HMM		Bits score	E-value	Alignment mode	Show/hide alignment
			Start	End	From	To				
RNA_ligase	RNA ligase	Family	25	407	1	443	859.2	2.1e-255	ls	Show

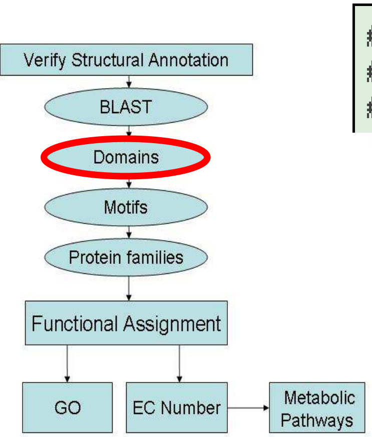
View the alignment:



```
#HMM       *->FkkYssLeNhyeskFIeklkmnGltggEWVArEKiHGaNFSliieedekeaqDGaeftVtyAKRsGiiGanvlPaE
#MATCH        F++Y++++N++e++ I++lk++G++++EW+A+EK+HGaNF+++++e+ek        +++yAKRsGi+    +P+E
#SEQ          FERYTEIDNSNERR-INALKGCGMFEDEWIATEKVHGANFGIYSIEGEK--------MIRYAKRSGIM----PPNE
```

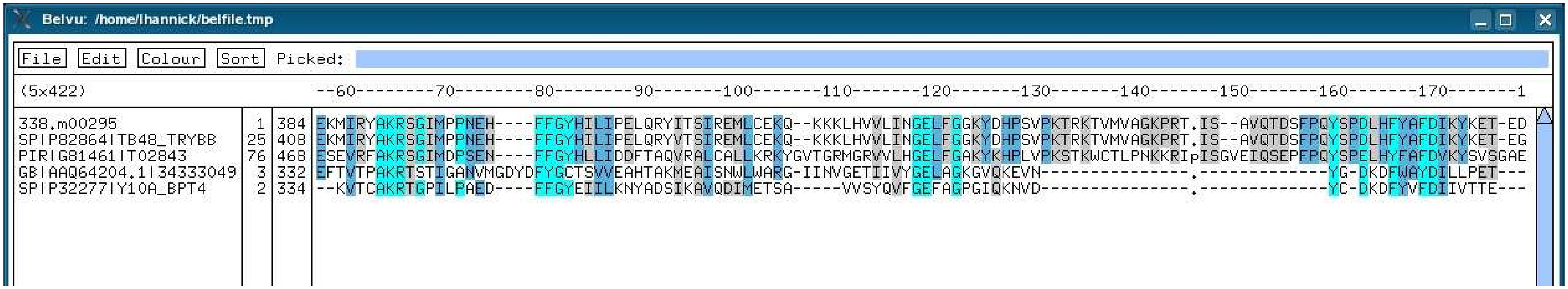


```
ldFyGYeivikdyaaaikavqelLetkqgvsGiyvrlevvqvyGELaGgKYdHPsVPKsrktvmvagkkriPrtivgvQkevFPdYgPDkl
E+F+GY+i+i++++++i++++e+L++kq+++     l+vv+++GEL+GgKYdHPsVPK+rktvmvagk   Prti++vQ+++FP+Y+PD+
EHFFGYHILIPELQRYITSIREMLCEKQKKK-----LHVVLINGELFGGKYDHPSVPKTRKTVMVAGK---PRTISAVQTDSFPQYSPDL
```



? See Glossary for HMM scores

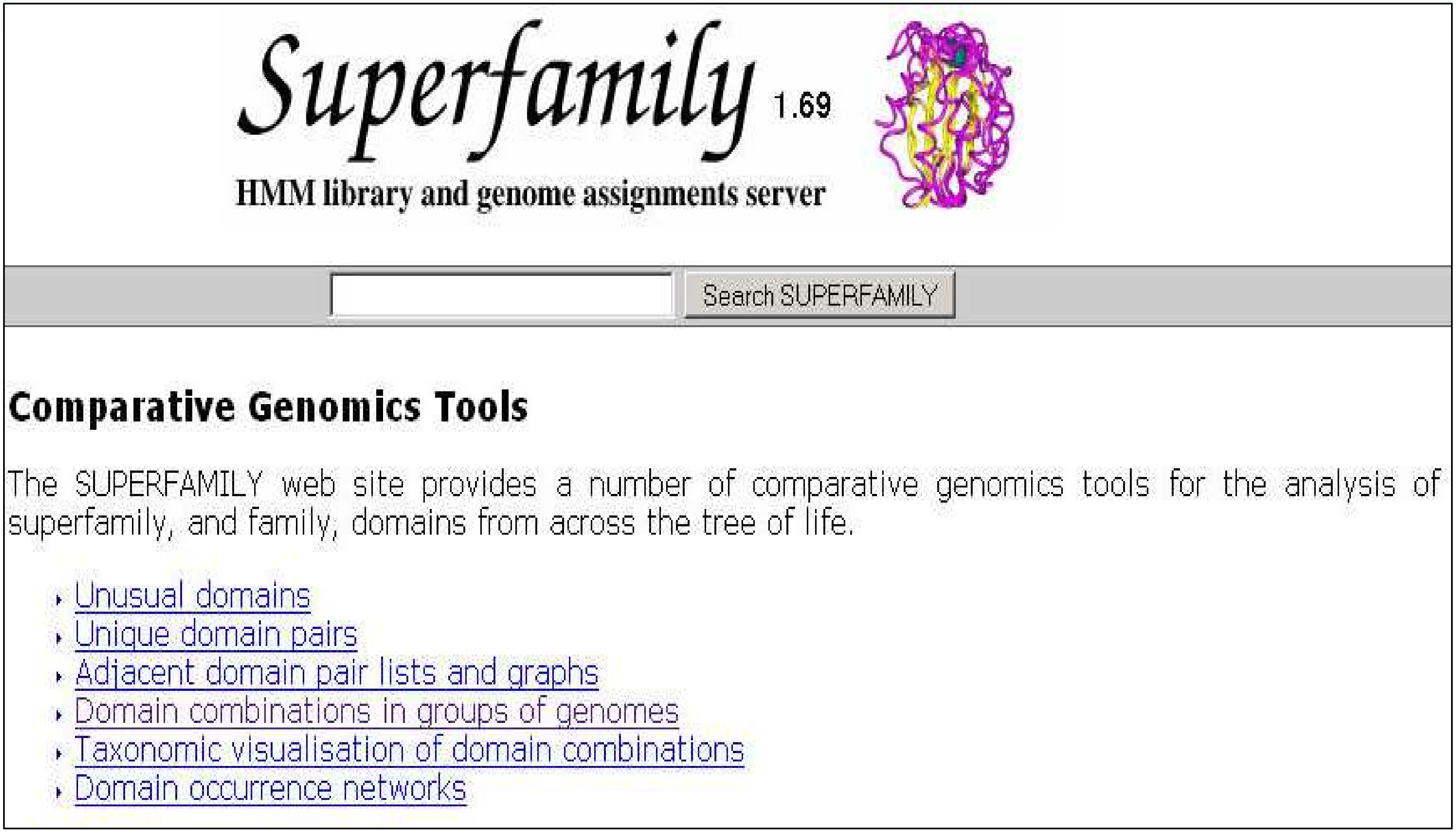
Verify HMM alignment



```
(5x422)                               --60--------70--------80--------90-------100-------110-------120-------130-------140-------150-------160-------170-------1
338.m00295                 1 384 EKMIRYAKRSGIMPPNEH----FFGYHILIPELQRYITSIREMLCEKQ--KKKLHVVLINGELFGGKYDHPSVPKTRKTVMVAGKPRT.IS--AVQTDSFPQYSPDLHFYAFDIKYKET-ED
SP|P82864|TB48_TRYBB      25 408 EKMIRYAKRSGIMPPNEH----FFGYHILIPELQRYVTSIREMLCEKQ--KKKLHVVLINGELFGGKYDHPSVPKTRKTVMVAGKPRT.IS--AVQTDSFPQYSPDLHFYAFDIKYKET-EG
PIR|G81461|T02843         76 468 ESEVRFAKRSGIMDPSEN----FFGYHLLIDDFTAQVRALCALLKRKYGVTGRMGRVVLHGELFGAKYKHPLVPKSTKWCTLPNKKRIpISGVEIQSEPFPQYSPELHYFAFDVKYSVSGAE
GB|AAQ64204.1|34333049     3 332 EFTVTPAKRTSTIGANVMGDYDFYGCTSVVEAHTAKMEAISNWLWARG-IINVGETIIVYGELAGKGVQKEVN---------------.-------------YG-DKDFWAYDILLPET---
SP|P32277|Y10A_BPT4        2 334 --KVTCAKRTGPILPAED----FFGYEIILKNYADSIKAVQDIMETSA-----VVSYQVFGEFAGPGIQKNVD---------------.-------------YC-DKDFYVFDIIVTTE---
```

Our sequence contains an RNA ligase, Rnl2 family domain, with a very strong match. Members of this Pfam family ligate (seal breaks in) RNA.

Superfamily



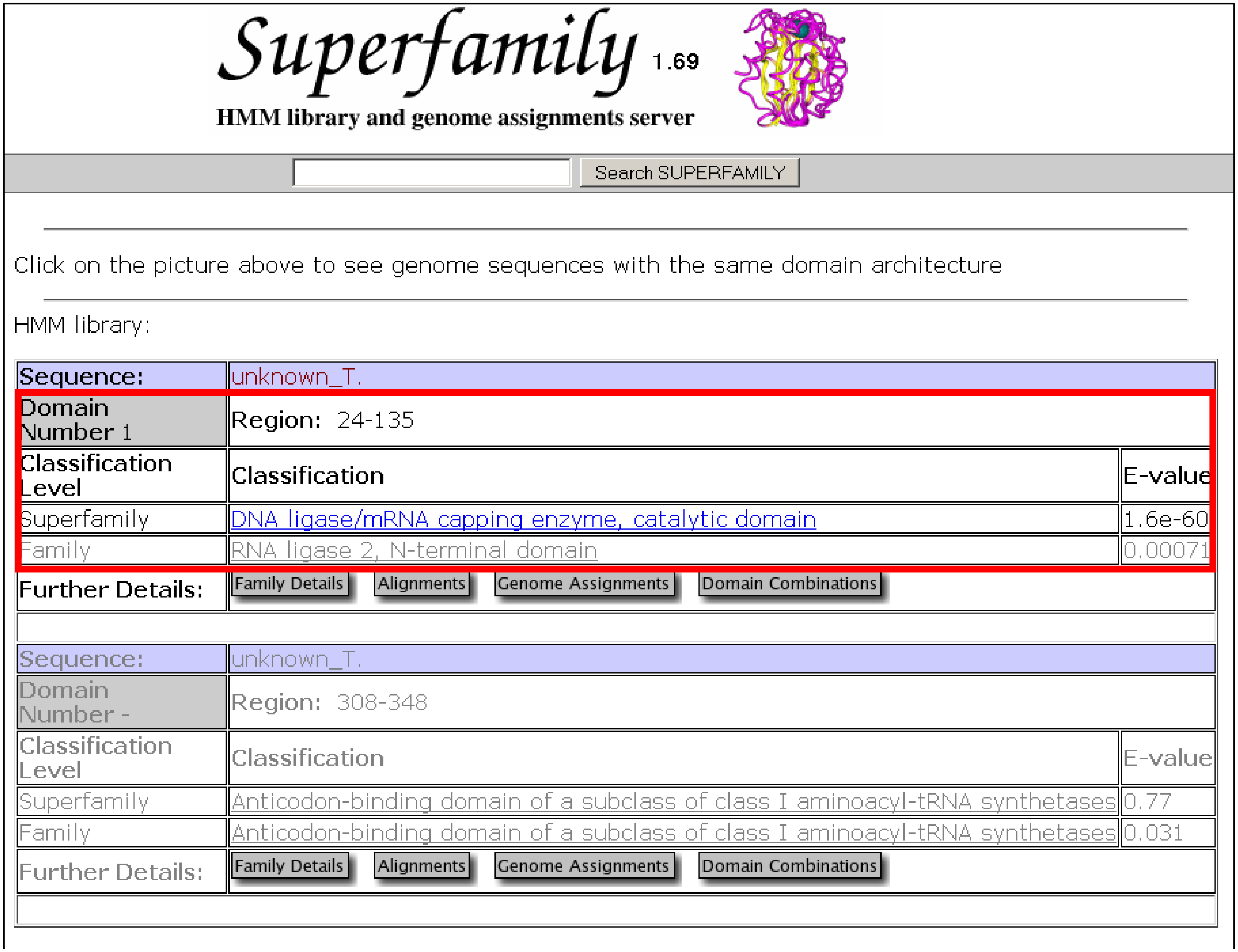
Comparative Genomics Tools

The SUPERFAMILY web site provides a number of comparative genomics tools for the analysis of superfamily, and family, domains from across the tree of life.

- Unusual domains
- Unique domain pairs
- Adjacent domain pair lists and graphs
- Domain combinations in groups of genomes
- Taxonomic visualisation of domain combinations
- Domain occurrence networks

Superfamily uses SCOP structural domains.

Superfamily Result



Click on the picture above to see genome sequences with the same domain architecture

HMM library:

Sequence:	unknown_T.	
Domain Number 1	Region: 24-135	
Classification Level	Classification	E-value
Superfamily	DNA ligase/mRNA capping enzyme, catalytic domain	1.6e-60
Family	RNA ligase 2, N-terminal domain	0.00071
Further Details:	Family Details Alignments Genome Assignments Domain Combinations	

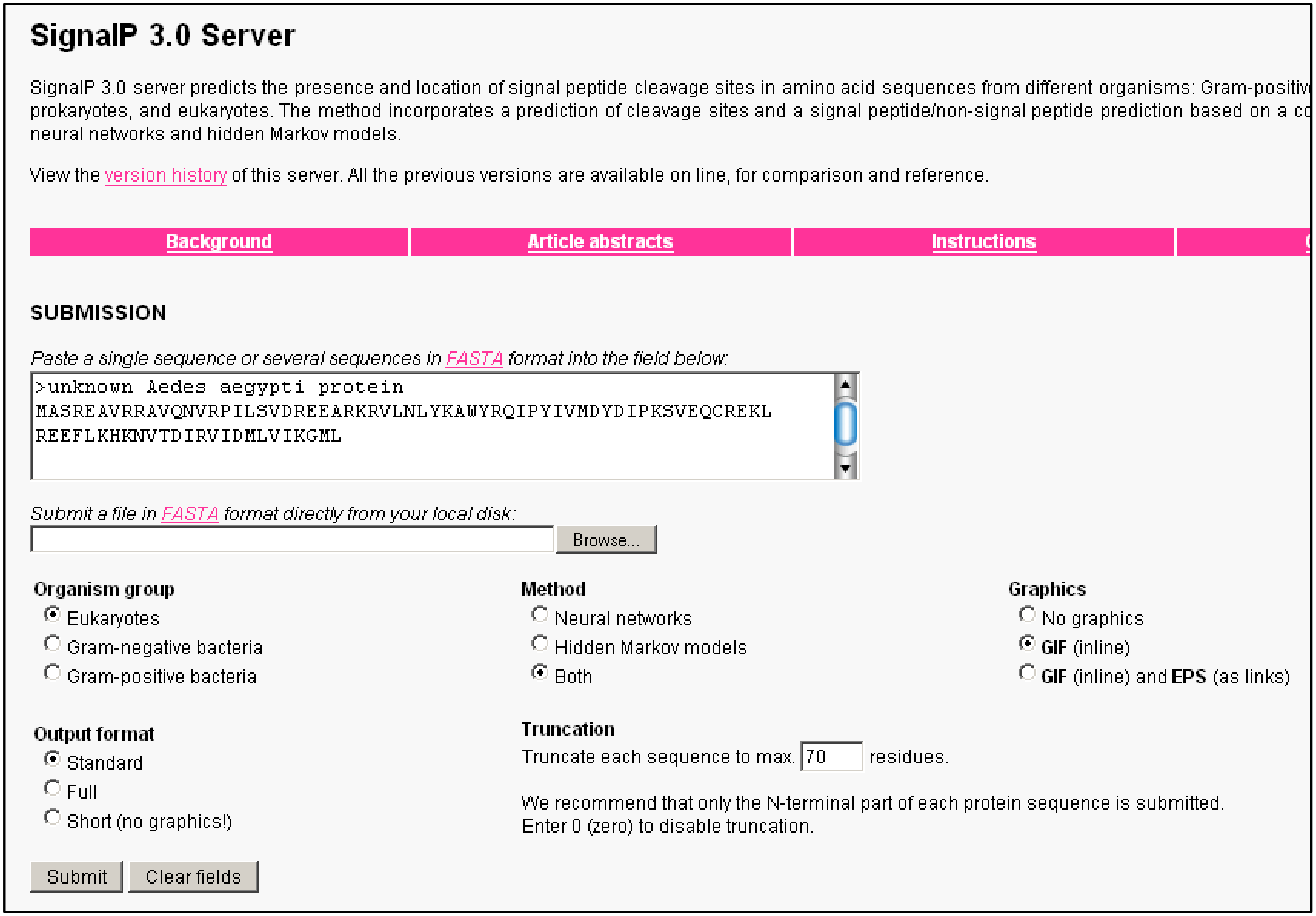
Sequence:	unknown_T.	
Domain Number -	Region: 308-348	
Classification Level	Classification	E-value
Superfamily	Anticodon-binding domain of a subclass of class I aminoacyl-tRNA synthetases	0.77
Family	Anticodon-binding domain of a subclass of class I aminoacyl-tRNA synthetases	0.031
Further Details:	Family Details Alignments Genome Assignments Domain Combinations	

SignalP

SignalP predicts the presence and location of signal peptide cleavage sites in amino acid sequences from different organisms: Gram-positive prokaryotes, Gram-negative prokaryotes, and eukaryotes.

http://www.cbs.dtu.dk/services/SignalP/

The method incorporates a prediction of cleavage sites and a signal peptide/non-signal peptide prediction based on a combination of several artificial neural networks and hidden Markov models.



SignalP 3.0 Server

SignalP 3.0 server predicts the presence and location of signal peptide cleavage sites in amino acid sequences from different organisms: Gram-positiv prokaryotes, and eukaryotes. The method incorporates a prediction of cleavage sites and a signal peptide/non-signal peptide prediction based on a co neural networks and hidden Markov models.

View the version history of this server. All the previous versions are available on line, for comparison and reference.

SUBMISSION

Paste a single sequence or several sequences in FASTA format into the field below:

```
>unknown Aedes aegypti protein
MASREAVRRAVQNVRPILSVDREEARKRVLNLYKAWYRQIPYIVMDYDIPKSVEQCREKL
REEFLKHKNVTDIRVIDMLVIKGML
```

Submit a file in FASTA format directly from your local disk:

Organism group
● Eukaryotes
○ Gram-negative bacteria
○ Gram-positive bacteria

Method
○ Neural networks
○ Hidden Markov models
● Both

Graphics
○ No graphics
● **GIF** (inline)
○ **GIF** (inline) and **EPS** (as links)

Output format
● Standard
○ Full
○ Short (no graphics!)

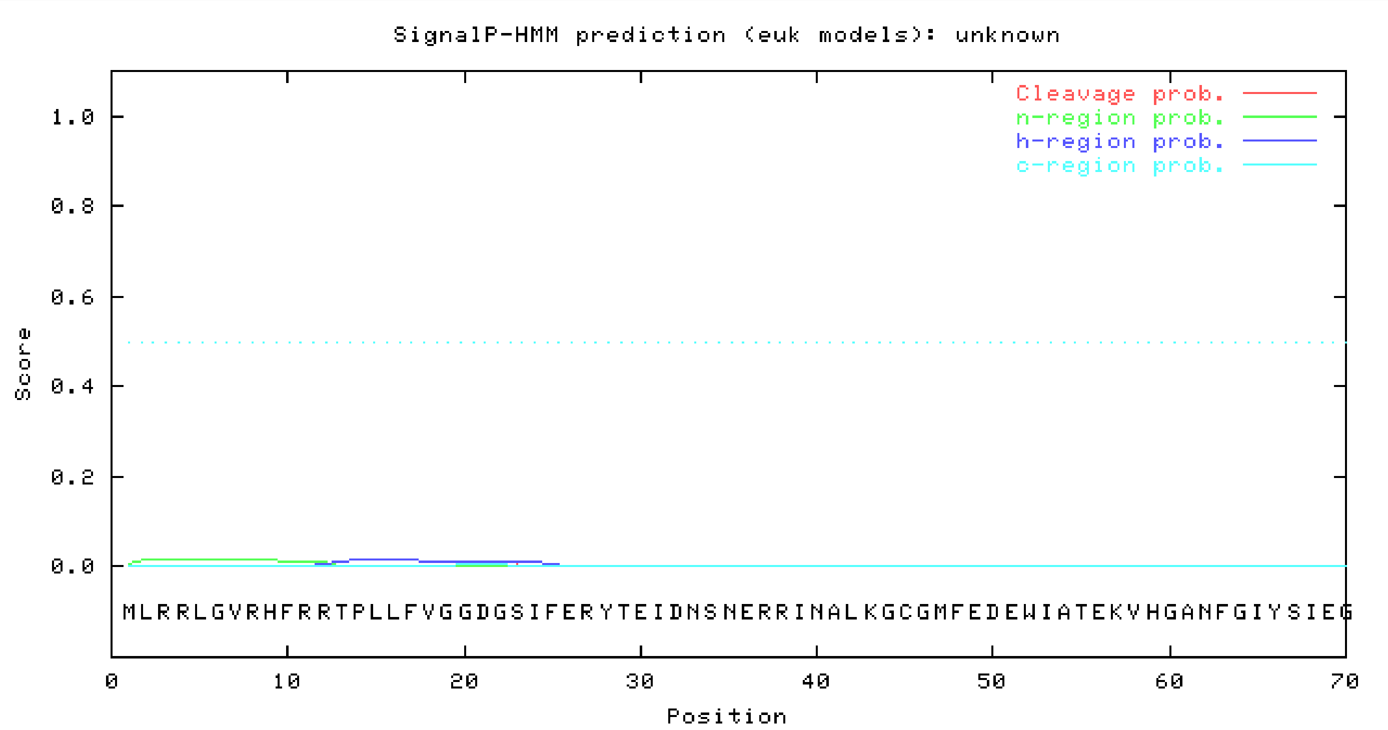
Truncation
Truncate each sequence to max. 70 residues.

We recommend that only the N-terminal part of each protein sequence is submitted.
Enter 0 (zero) to disable truncation.

SignalP results

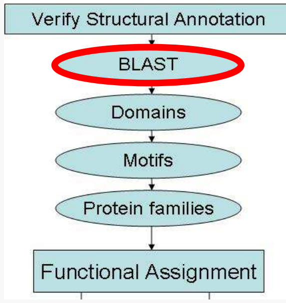
Non-secretory protein

SignalP-HMM result:



data

```
>unknown
Prediction: Non-secretory protein
Signal peptide probability: 0.008
Signal anchor probability: 0.009
Max cleavage site probability: 0.006 between pos. 22 and 23
```



See Glossary entry for Signal Peptide

TargetP

http://www.cbs.dtu.dk/services/TargetP/

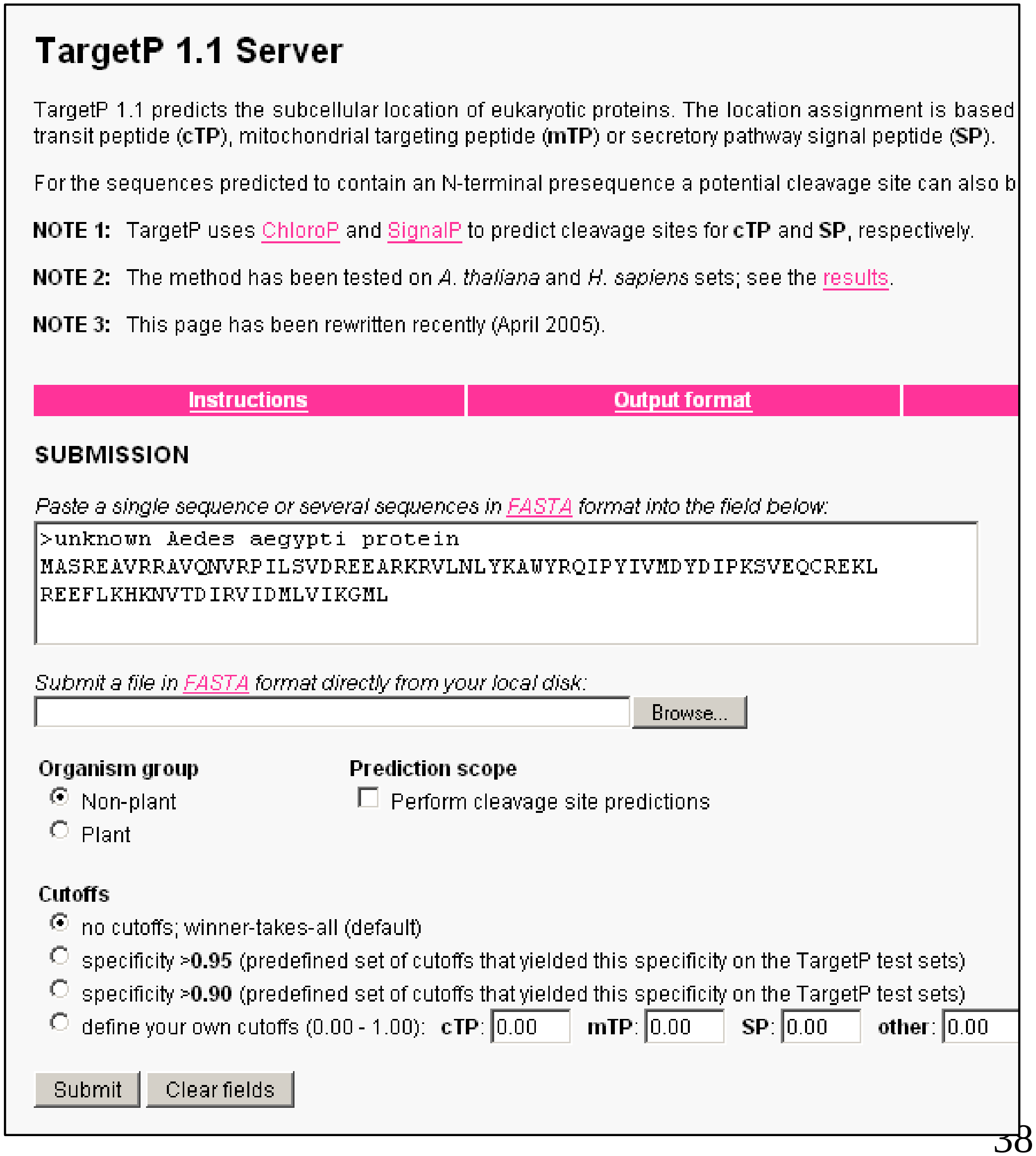
TargetP predicts the subcellular location of eukaryotic proteins.

The location assignment is based on the predicted presence of any of the N-terminal presequences:

➢chloroplast transit peptide (**cTP)**

➢mitochondrial targeting peptide **(mTP)**

➢secretory pathway signal peptide **(SP).**



TargetP 1.1 Server

TargetP 1.1 predicts the subcellular location of eukaryotic proteins. The location assignment is based transit peptide (**cTP**), mitochondrial targeting peptide (**mTP**) or secretory pathway signal peptide (**SP**).

For the sequences predicted to contain an N-terminal presequence a potential cleavage site can also b

NOTE 1: TargetP uses ChloroP and SignalP to predict cleavage sites for **cTP** and **SP**, respectively.

NOTE 2: The method has been tested on *A. thaliana* and *H. sapiens* sets; see the results.

NOTE 3: This page has been rewritten recently (April 2005).

SUBMISSION

Paste a single sequence or several sequences in FASTA format into the field below:

```
>unknown Aedes aegypti protein
MASREAVRRAVQNVRPILSVDREEARKRVLNLYKAWYRQIPYIVMDYDIPKSVEQCREKL
REEFLKHKNVTDIRVIDMLVIKGML
```

Submit a file in FASTA format directly from your local disk:

Organism group

● Non-plant

○ Plant

Prediction scope

□ Perform cleavage site predictions

Cutoffs

● no cutoffs; winner-takes-all (default)

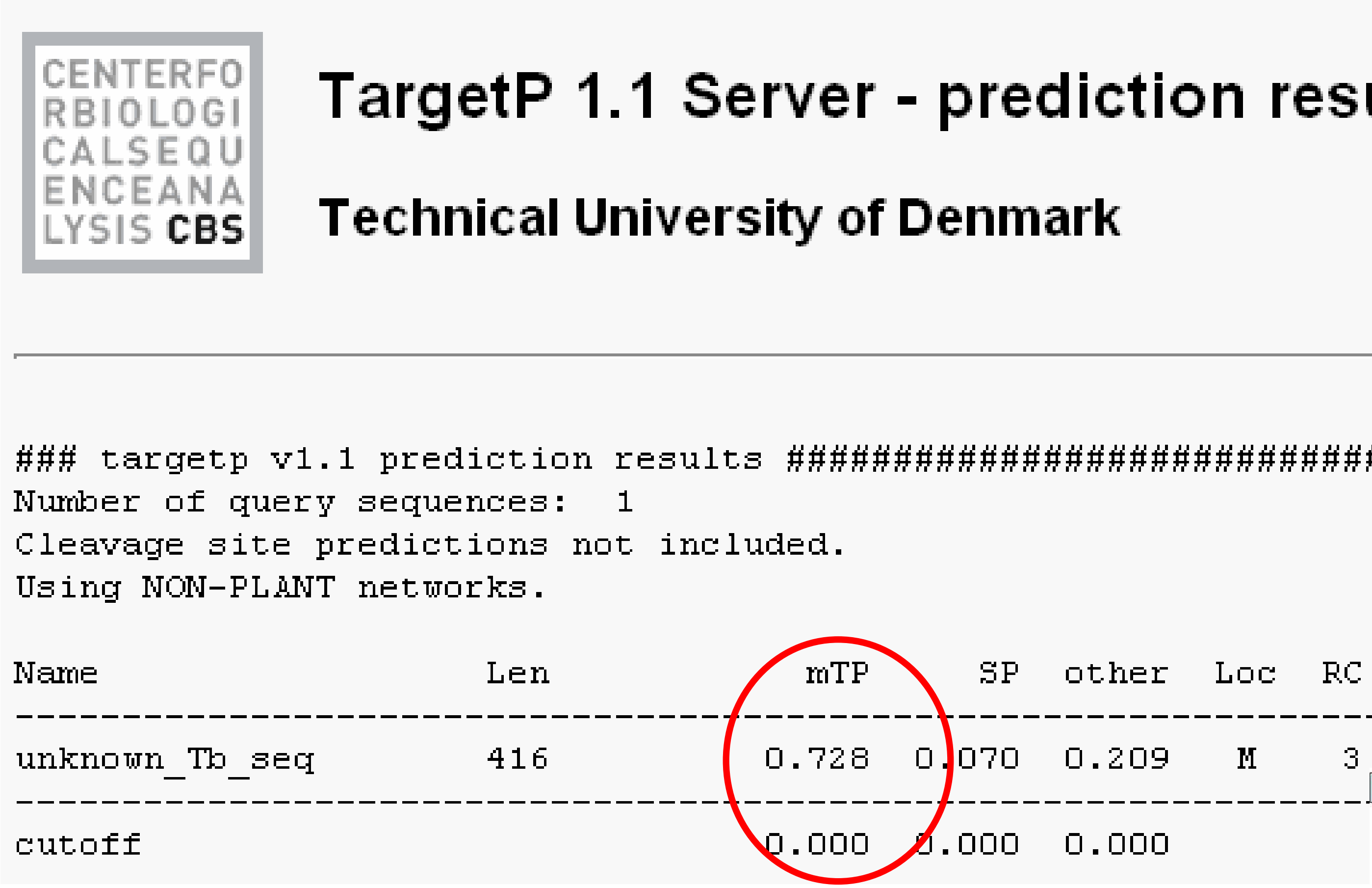
○ specificity >**0.95** (predefined set of cutoffs that yielded this specificity on the TargetP test sets)

○ specificity >**0.90** (predefined set of cutoffs that yielded this specificity on the TargetP test sets)

○ define your own cutoffs (0.00 - 1.00): **cTP**: 0.00 **mTP**: 0.00 **SP**: 0.00 **other**: 0.00

TargetP results

The sequence contains a mitochondrial targeting peptide, mTP.

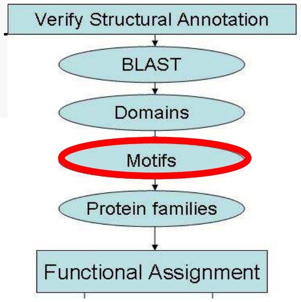


CENTERFORBIOLOGICALSEQUENCEANALYSIS CBS

TargetP 1.1 Server - prediction results

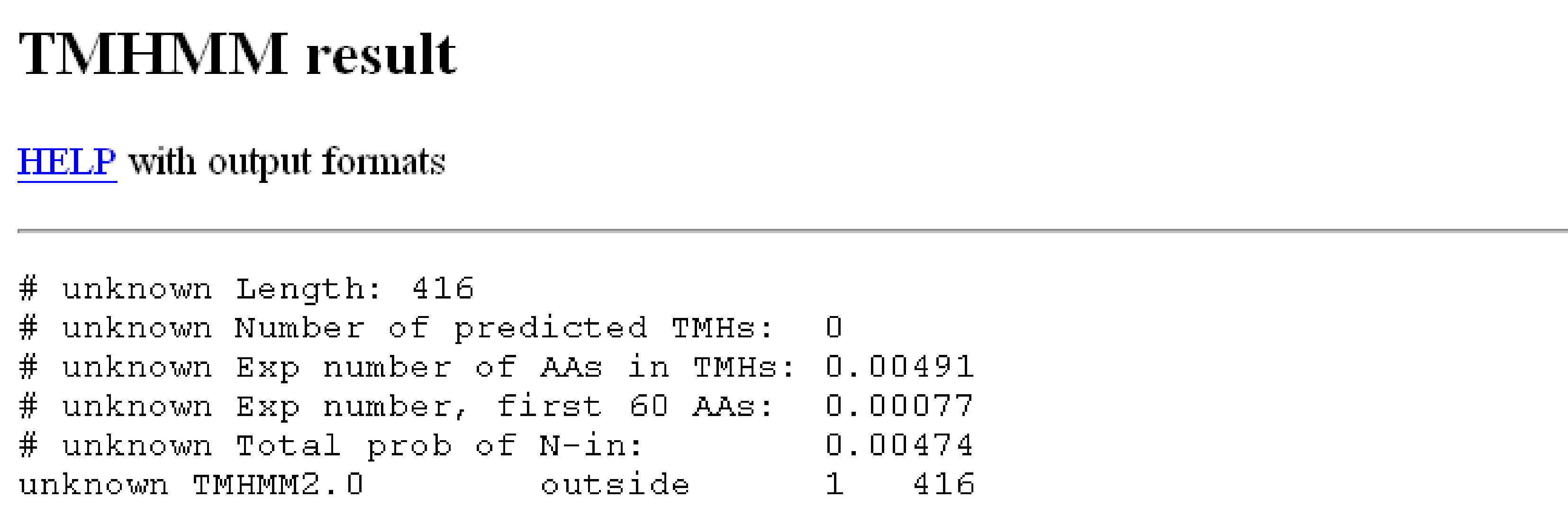
Technical University of Denmark

```
### targetp v1.1 prediction results ##################################
Number of query sequences:  1
Cleavage site predictions not included.
Using NON-PLANT networks.

Name                  Len            mTP     SP  other  Loc  RC
----------------------------------------------------------------------
unknown_Tb_seq        416          0.728  0.070  0.209   M    3
----------------------------------------------------------------------
cutoff                             0.000  0.000  0.000
```



? See explanation of TargetP output in extra slides.

Transmembrane domains

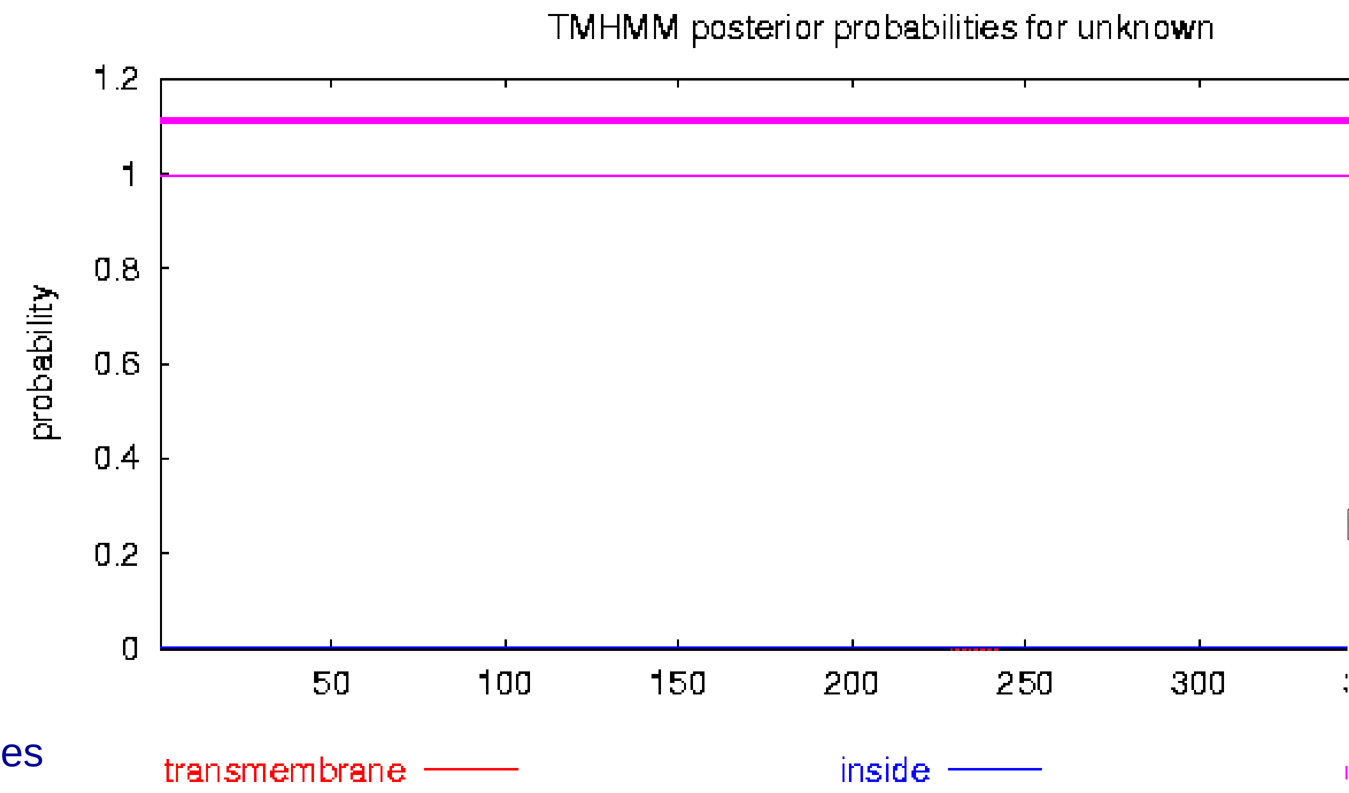


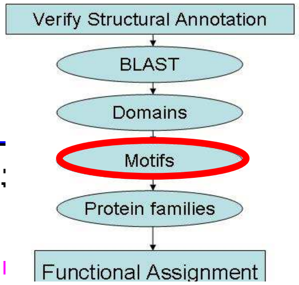
TMHMM result

HELP with output formats

```
# unknown Length: 416
# unknown Number of predicted TMHs:  0
# unknown Exp number of AAs in TMHs: 0.00491
# unknown Exp number, first 60 AAs:  0.00077
# unknown Total prob of N-in:        0.00474
unknown TMHMM2.0        outside      1   416
```

There are no transmembrane domains.





? See slide 58 for explanation of scores

Annotation of Example 1

BLAST: A protein match at Swiss-Prot is 99% identical, with 2 conservative and one non-conservative amino acid substitutions. “RNA-editing ligase TbMP48, mitochondrial precursor” is the Swiss-Prot name for this close protein match.

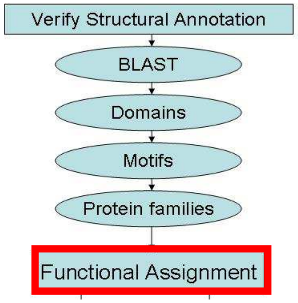
This mitochondrial precursor of an RNA ligase was identified as a member of a multi-protein complex that catalyzes deletion editing in vitro. It was isolated from an enriched sample of Trypanosoma brucei mitochondria by sequential ion-exchange and gel filtration chromatography, followed by glycerol gradient sedimentation. The protein was not functionally characterized, but was identified as a member of an RNA-editing complex. The complex was shown to have RNA-editing function. (PMID:11134327)

Domain: Our sequence contains an RNA ligase, Rnl2 family, with a very strong match. Members of this family ligate (seal breaks in) RNA.

Signal sequence: none

Targeting Sequence: It contains a mitochondrial targeting sequence.

Under the standards of the Tri-tryp project, “RNA-editing ligase TbMP48, mitochondrial precursor,” is a suitable name.



Evidence from homology searching

Compare sequences of unknown function to those of known function.

Shared sequence identity may imply shared function.

- Full-length match with significant identity (>35%)
- Domains and motifs
- Binding sites
- Catalytic sites

But beware

- there are occurrences where one amino acid substitution changes the function of an enzyme.
- synonymous or “silent” codon substitutions may result in functional differences.*
- Mutations may result in modification or deletion of function.
- all functional assignments made by similarity should be considered tentative until confirmed by experiment.

***** Kimchi-Sarfaty C, Oh JM, Kim IW, Sauna ZE, Calcagno AM, Ambudkar SV, Gottesman MM. A "silent" polymorphism in the MDR1 gene changes substrate specificity**.** *Science* 2007 Jan 26 **315**(5811):525-8

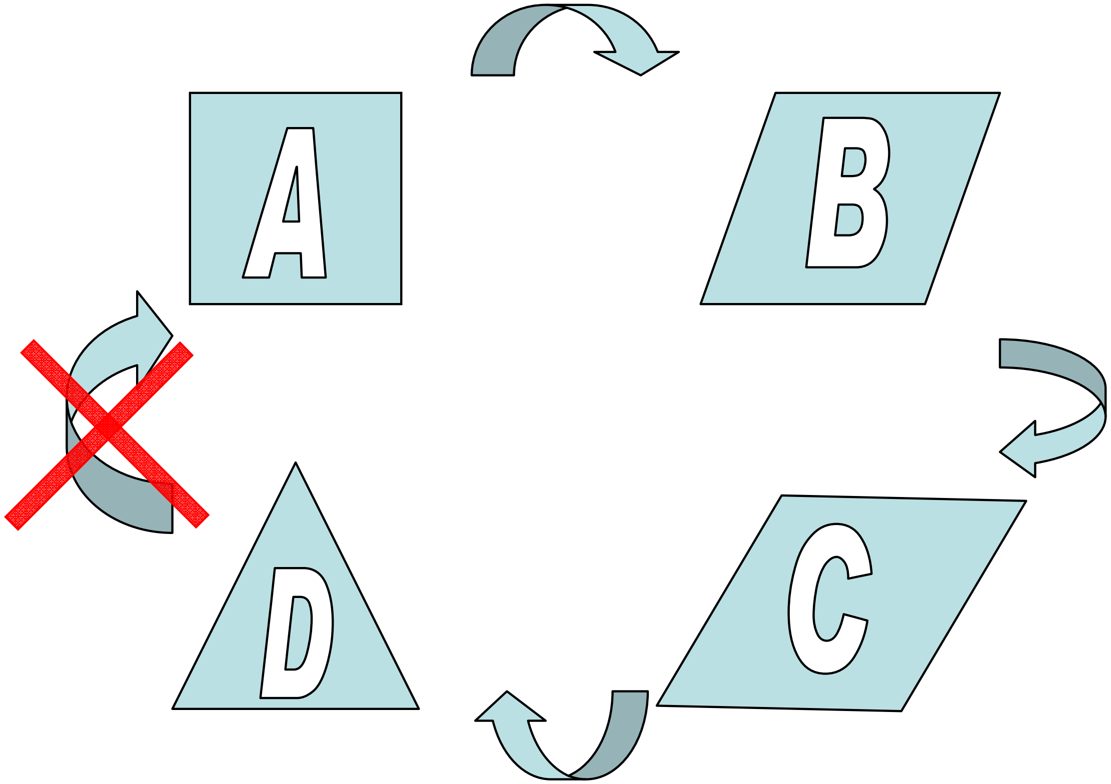
Transitive annotation

A is like B

B is like C

C is like D

D is NOT like A!



Take a conservative approach. Err on the side of missing homology rather than stretching weak data.

Not experimentally characterized…

The fun begins when you need to draw conclusions about genes and gene products that have not been characterized.

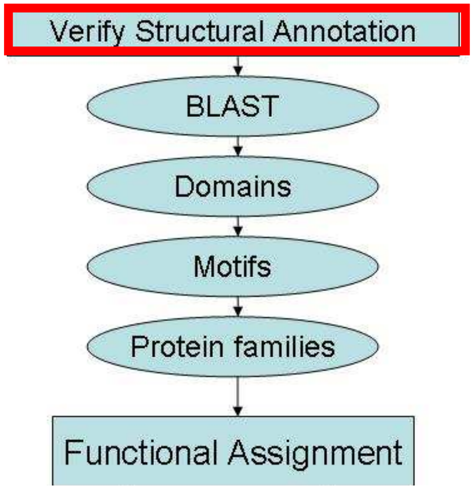
Examine all possible sources of information!

- If you have automated annotation results, verify them.
- HMM: is/are the domain hit(s) significant?
- Is there a signal sequence, a targeting sequence?
- Does it belong to a family of proteins or genes?
- What do the homology searches tell you?

Example 2

Our second example is an unknown Aedes aegypti protein sequence

```
>unknown_Aedes_aegypti_protein_85aa
MASREAVRRAVQNVRPILSVDREEARKRVLN
LYKAWYRQIPYIVMDYDIPKSVEQCREKLRE
EFLKHKNVTDIRVIDMLVIKGML
```



Verify Structural Annotation

BLAST

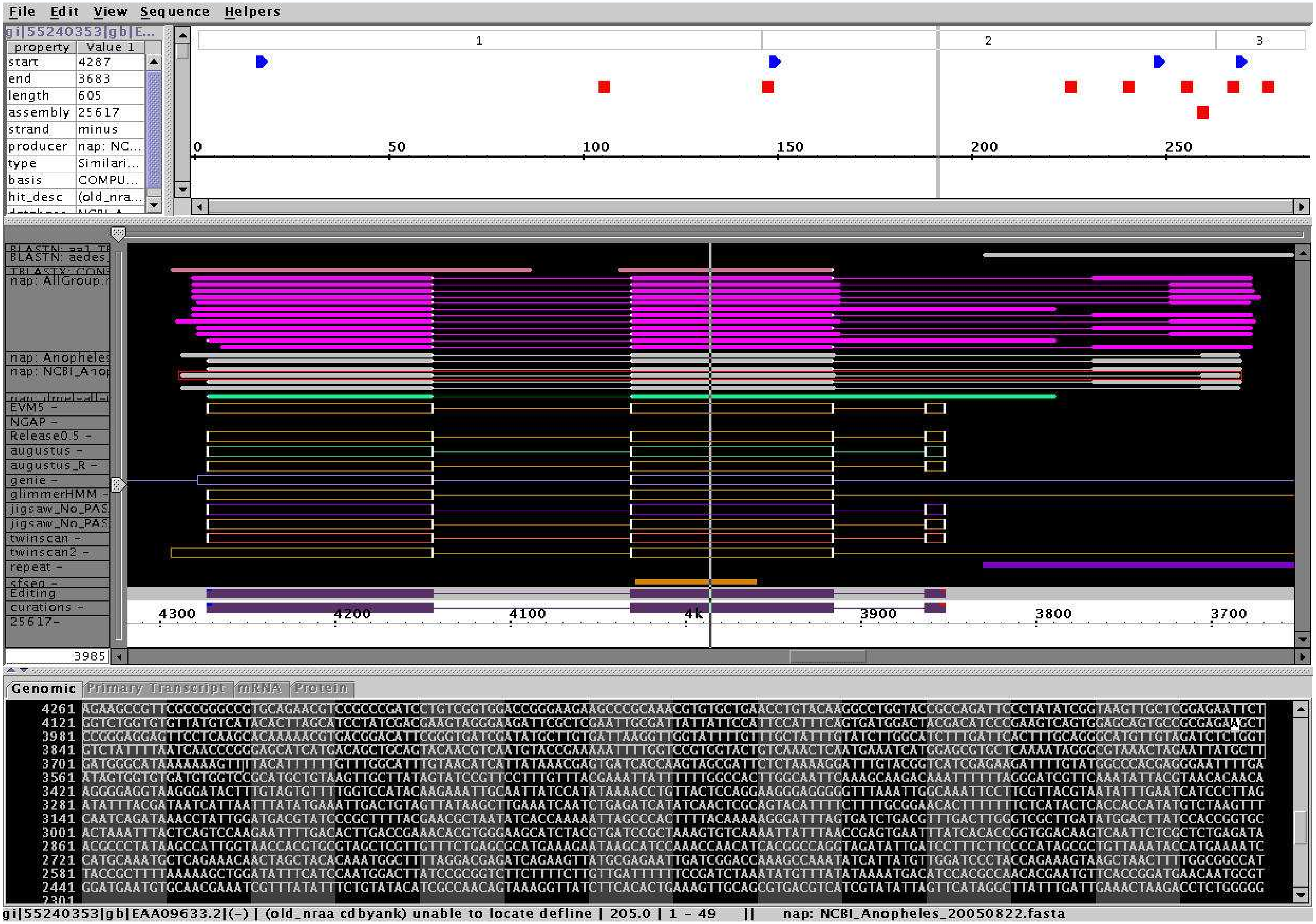
Domains

Motifs

Protein families

Functional Assignment

Example 2: verify gene structure



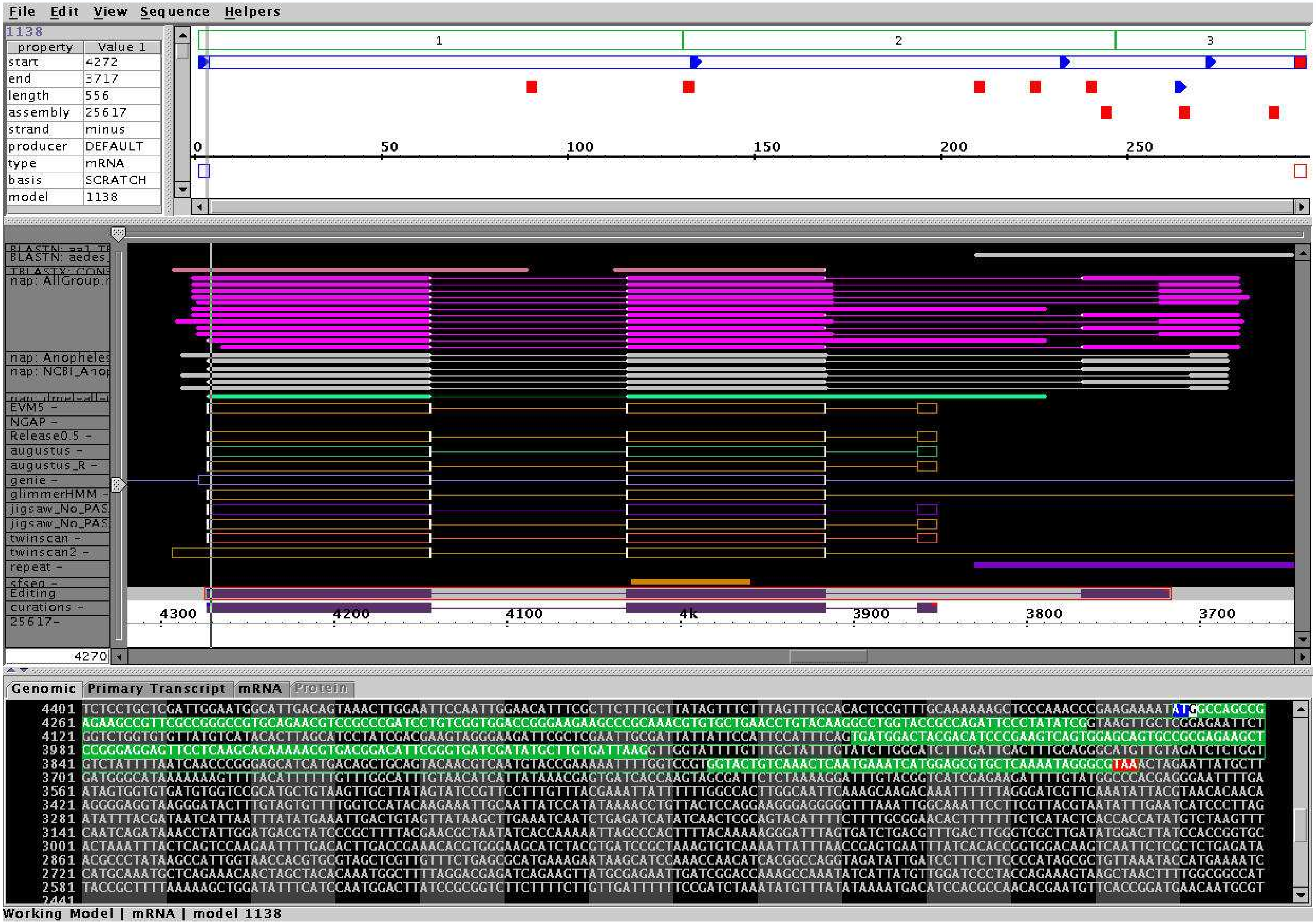
gi|55240353|gb|E...

property	Value 1
start	4287
end	3683
length	605
assembly	25617
strand	minus
producer	nap: NC...
type	Similari...
basis	COMPU...
hit_desc	(old_nra...
databas...	NCBI A...

```
4261 AGAAGCCGTTCGCCGGGCCGTGCAGAACGTCCGCCCCGATCCTGTCGGTGGACCGGGAAGAAGCCCCGCAAACGTGTGCTGAACCTGTACAAGGCCTGGTACCGCCAGATTCCCTATATCGGTAAGTTGCTCGGAGAATTCT
4121 GGTCTGGTGTGTTATGTCATACACTTAGCATCCTATCGACGAAGTAGGGAAGATTCGCTCGAATTGCGATTATTATTCCATTCCATTTCAGTGATGGACTACGACATCCCGAAGTCAGTGGAGCAGTGCCGCGAGAAGCT
3981 CCGGGAGGAGTTCCTCAAGCACAAAAACGTGACGGACATTCGGGTGATCGATATGCTTGTGATTAAGGTTGGTATTTTGTTTGCTATTTGTATCTTGGCATCTTTGATTCACTTTGCAGGGCATGTTGTAGATCTCTGGT
3841 GTCTATTTTAATCAACCCGGGAGCATCATGACAGCTGCAGTACAACGTCAATGTACCGAAAAATTTTGGTCCGTGGTACTGTCAAACTCAATGAAATCATGGAGCGTGCTCAAAATAGGGCGTAAACTAGAATTATGCTT
3701 GATGGCCATAAAAAAAGTTTTACATTTTTTGTTTGGCATTTGTAACATCATTATAAACGAGTGATCACCAAGTAGCGATTCTCTAAAAGGATTTGTACGGTCATCGAGAAGATTTTGTATGGCCCACGAGGGAATTTTGA
3561 ATAGTGGTGTGATGTGGTCCGCATGCTGTAAGTTGCTTATAGTATCCGTTCCTTTGTTTACGAAATTATTTTTTGGCCACTTGGCAATTCAAAGCAAGACAAATTTTTTAGGGATCGTTCAAATATTACGTAACACAACA
3421 AGGGGAGGTAAGGGATACTTTGTAGTGTTTTGGTCCATACAAGAAATTGCAATTATCCATATAAAACCTGTTACTCCAGGAAGGGAGGGGGTTTAAATTGGCAAATTCCTTCGTTACGTAATATTTGAATCATCCCTTAG
3281 ATATTTACGATAATCATTAATTTATATGAAATTGACTGTAGTTATAAGCTTGAAATCAATCTGAGATCATATCAACTCGCAGTACATTTTCTTTTGCGGAACACTTTTTTTCTCATACTCACCACCATATGTCTAAGTTT
3141 CAATCAGATAAACCTATTGGATGACGTATCCCGCTTTTACGAACGCTAATATCACCAAAAATTAGCCCACTTTTACAAAAAGGGATTTAGTGATCTGACGTTTGACTTGGGTCGCTTGATATGGACTTATCCACCGGTGC
3001 ACTAAATTTACTCAGTCCAAGAATTTTGACACTTGACCGAAACACGTGGGAAGCATCTACGTGATCCGCTAAAGTGTCAAAATTATTTAACCGAGTGAATTTATCACACCGGTGGACAAGTCAATTCTCGCTCTGAGATA
2861 ACGCCCTATAAGCCATTGGTAACCACGTGCGTAGCTCGTTGTTTCTGAGCGCATGAAAGAATAAGCATCCAAACCAACATCACGGCCAGGTAGATATTGATCCTTTCTTCCCCATAGCGCTGTTAAATACCATGAAAATC
2721 CATGCAAATGCTCAGAAACAACTAGCTACACAAATGGCTTTTAGGACGAGATCAGAAGTTATGCGAGAATTGATCGGACCAAAGCCAAATATCATTATGTTGGATCCCTACCAGAAAGTAAGCTAACTTTTGGCGGCCAT
2581 TACCGCTTTTAAAAAGCTGGATATTTCATCCAATGGACTTATCCGCGGTCTTCTTTTCTTGTTGATTTTTTCCGATCTAAATATGTTTATATAAAATGACATCCACGCCAACACGAATGTTCACCGGATGAACAATGCGT
2441 GGATGAATGTGCAACGAAATCGTTTATATTTCTGTATACATCGCCAACAGTAAAGGTTATCTTCACACTGAAAGTTGCAGCGTGACGTCATCGTATATTAGTTCATAGGCTTATTTGATTGAAACTAAGACCTCTGGGG
2301
```

gi|55240353|gb|EAA09633.2|(-) | (old_nraa cdbyank) unable to locate defline | 205.0 | 1 - 49 || nap: NCBI_Anopheles_20050822.fasta

Correct the gene structure



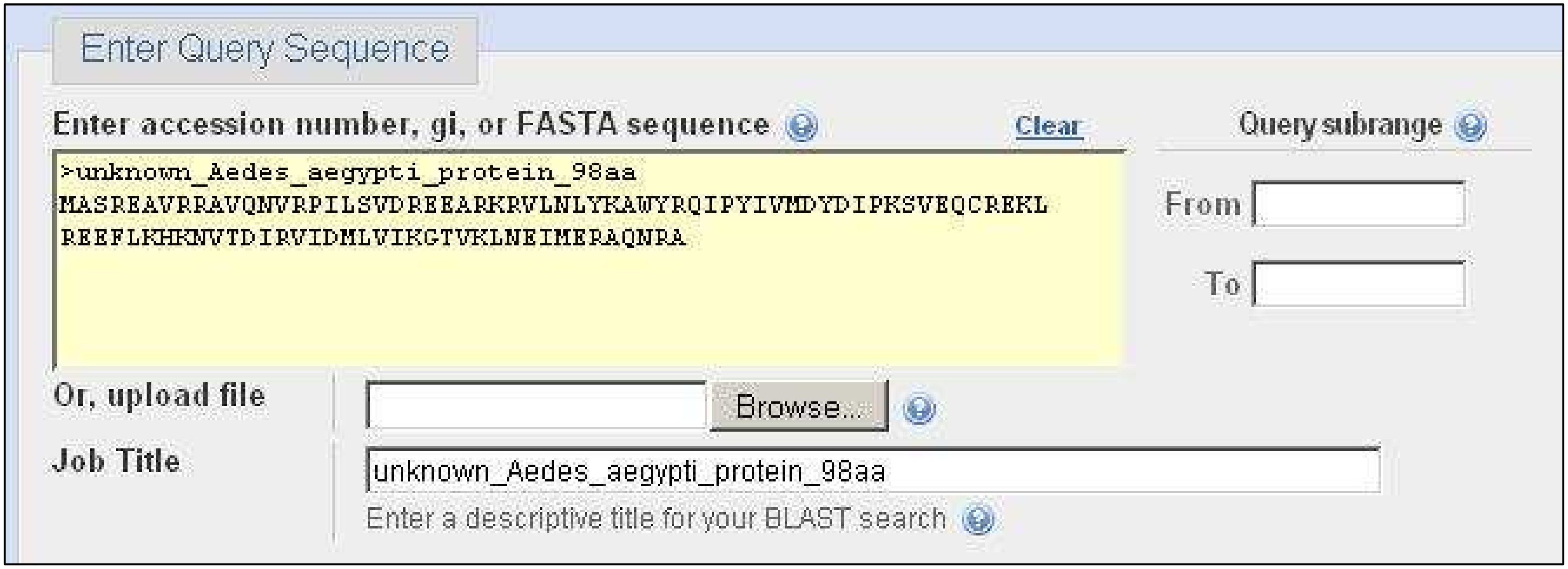
property	Value 1
start	4272
end	3717
length	556
assembly	25617
strand	minus
producer	DEFAULT
type	mRNA
basis	SCRATCH
model	1138

```
4401 TCTCCTGCTCGATTGGAATGGCATTGACAGTAAACTTGGAATTCCAATTGGAACATTTCGCTTCTTTGCTTATAGTTTCTTTAGTTTGCACACTCCGTTTGCAAAAAAGCTCCCAAACCCGAAGAAAATATGGCCAGCCG
4261 AGAAGCCGTTCGCCGGGCCGTGCAGAACGTCCGCCCGATCCTGTCGGTGGACCGGGAAGAAGCCCGCAAACGTGTGCTGAACCTGTACAAGGCCTGGTACCGCCAGATTCCCTATATCGGTAAGTTGCTCGGAGAATTCT
4121 GGTCTGGTGTGTTATGTCATACACTTAGCATCCTATCGACGAAGTAGGGAAGATTCGCTCGAATTGCGATTATTATTCCATTCCATTTCAGTGATGGACTACGACATCCCGAAGTCAGTGGAGCAGTGCCGCGAGAAGCT
3981 CCGGGAGGAGTTCCTCAAGCACAAAAACGTGACGGACATTCGGGTGATCGATATGCTTGTGATTAAGGTTGGTATTTTGTTTGCTATTTGTATCTTGGCATCTTTGATTCACTTTGCAGGGCATGTTGTAGATCTCTGGT
3841 GTCTATTTTAATCAACCCGGGAGCATCATGACAGCTGCAGTACAACGTCAATGTACCGAAAAATTTTGGTCCGTGGTACTGTCAAACTCAATGAAATCATGGAGCGTGCTCAAAATAGGGCGTAAACTAGAATTATGCTT
3701 GATGGGCATAAAAAAAGTTTTACATTTTTTGTTTGGCATTTGTAACATCATTATAAACGAGTGATCACCAAGTAGCGATTCTCTAAAAGGATTTGTACGGTCATCGAGAAGATTTTGTATGGCCCACGAGGGAATTTTGA
3561 ATAGTGGTGTGATGTGGTCCGCATGCTGTAAGTTGCTTATAGTATCCGTTCCTTTGTTTACGAAATTATTTTTTTGGCCACTTGGCAATTCAAAGCAAGACAAATTTTTTTAGGGATCGTTCAAATATTACGTAACACAACA
3421 AGGGGAGGTAAGGGATACTTTGTAGTGTTTTGGTCCATACAAGAAATTGCAATTATCCATATAAAACCTGTTACTCCAGGAAGGGAGGGGGTTTAAATTGGCAAATTCCTTCGTTACGTAATATTTGAATCATCCCTTAG
3281 ATATTTACGATAATCATTAATTTATATGAAATTGACTGTAGTTATAAGCTTGAAATCAATCTGAGATCATATCAACTCGCAGTACATTTTCTTTTGCGGAACACTTTTTTTTCTCATACTCACCACCATATGTCTAAGTTT
3141 CAATCAGATAAACCTATTGGATGACGTATCCCGCTTTTACGAACGCTAATATCACCAAAAATTAGCCCACTTTTACAAAAAGGGATTTAGTGATCTGACGTTTGACTTGGGTCGCTTGATATGGACTTATCCACCGGTGC
3001 ACTAAATTTACTCAGTCCAAGAATTTTGACACTTGACCGAAACACGTGGGAAGCATCTACGTGATCCGCTAAAGTGTCAAAATTATTTAACCGAGTGAATTTATCACACCGGTGGACAAGTCAATTCTCGCTCTGAGATA
2861 ACGCCCTATAAGCCATTGGTAACCACGTGCGTAGCTCGTTGTTTCTGAGCGCATGAAAGAATAAGCATCCAAACCAACATCACGGCCAGGTAGATATTGATCCTTTCTTCCCCATAGCGCTGTTAAATACCATGAAAATC
2721 CATGCAAATGCTCAGAAACAACTAGCTACACAAATGGCTTTTAGGACGAGATCAGAAGTTATGCGAGAATTGATCGGACCAAAGCCAAATATCATTATGTTGGATCCCTACCAGAAAGTAAGCTAACTTTTGGCGGCCAT
2581 TACCGCTTTTAAAAAGCTGGATATTTCATCCAATGGACTTATCCGCGGTCTTCTTTTCTTGTTGATTTTTTCCGATCTAAATATGTTTATATAAAATGACATCCACGCCAACACGAATGTTCACCGGATGAACAATGCGT
2441
```

Working Model | mRNA | model 1138

BLASTP

```
>unknown_Aedes_aegypti_protein_98aa
MASREAVRRAVQNVRPILSVDREEARKRNLYKAWYRQIPYIVMDYDIPKSVE
QCREKLREEFLKHKNVTDIRVIDMLVIKGTVKLNEIMERAQNRA
```



Enter accession number, gi, or FASTA sequence

```
>unknown_Aedes_aegypti_protein_98aa
MASREAVRRAVQNVRPILSVDREEARKRVLNLYKAWYRQIPYIVMDYDIPKSVEQCREKL
REEFLKHKNVTDIRVIDMLVIKGTVKLNEIMERAQNRA
```

Query subrange

From

To

Or, upload file

Job Title: unknown_Aedes_aegypti_protein_98aa

Enter a descriptive title for your BLAST search

NCBI BLAST Results:

The first match is to itself

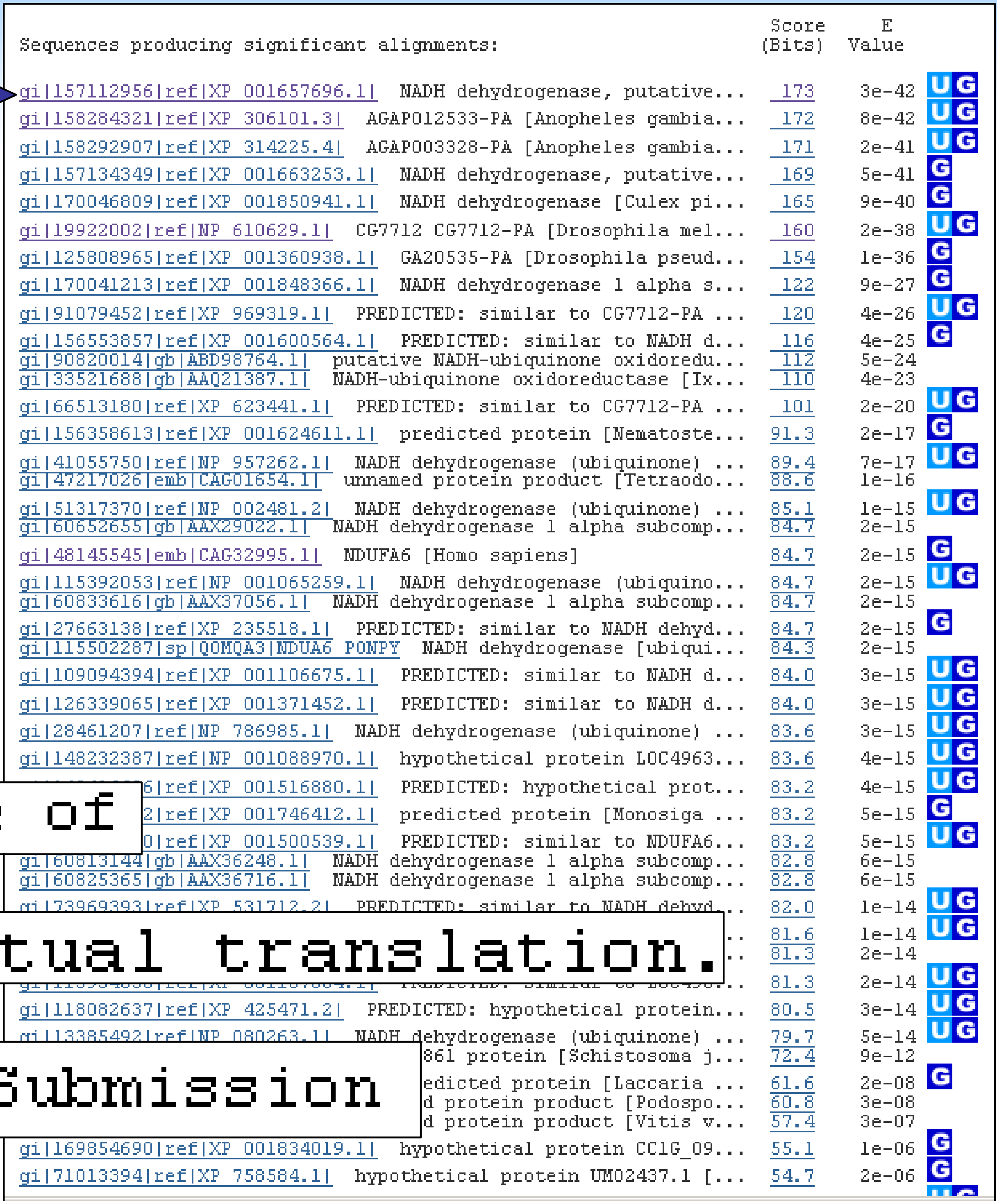
There are no significant blast hits to characterized proteins in the next 17 hits.

Some clues in the Genbank record that the entry is not characterized:

Genome Sequence of

Method: conceptual translation.

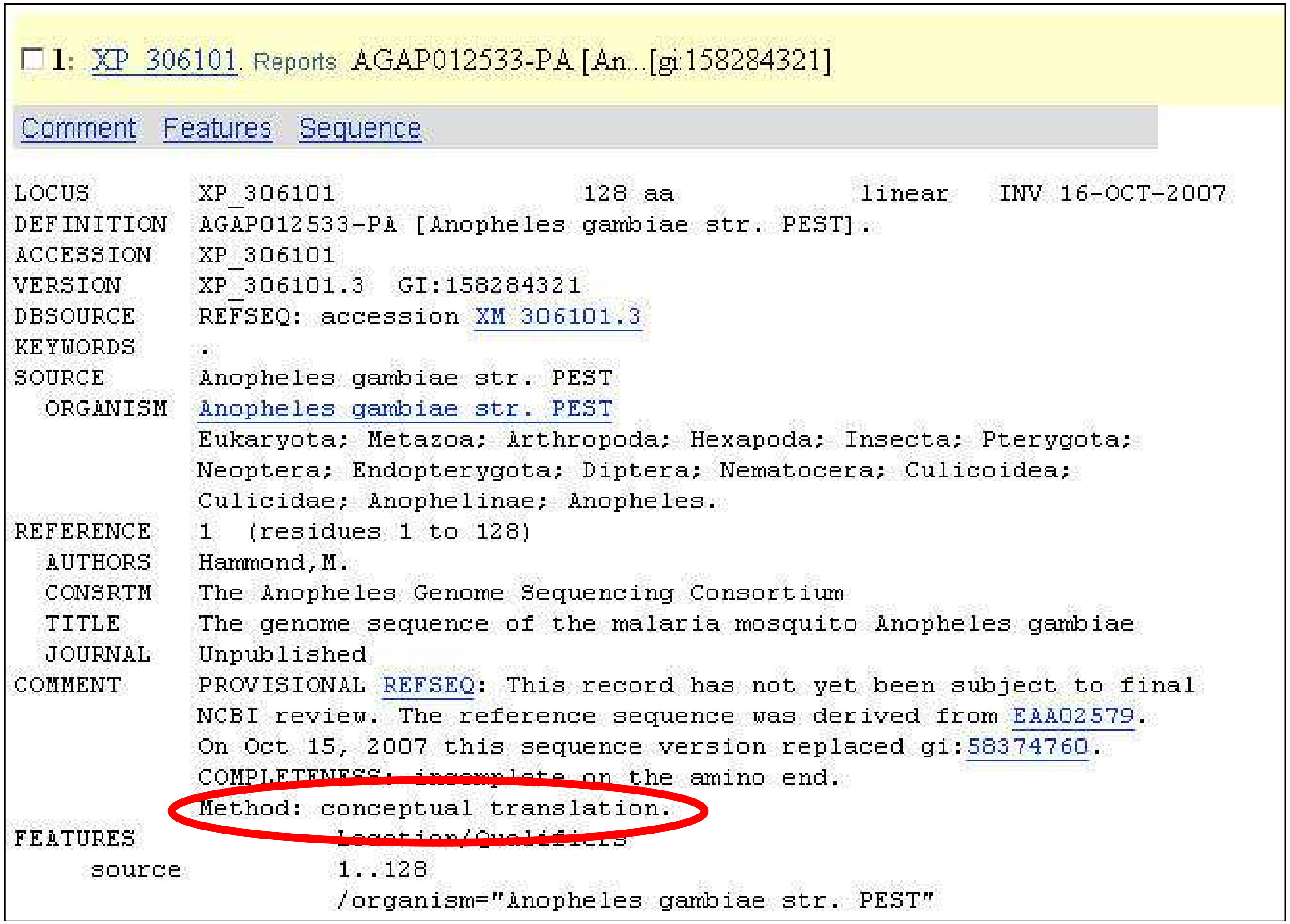
Direct Submission



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

gi|157112956|ref|XP 001657696.1|  NADH dehydrogenase, putative...  173    3e-42  UG
gi|158284321|ref|XP 306101.3|  AGAP012533-PA [Anopheles gambia...  172    8e-42  UG
gi|158292907|ref|XP 314225.4|  AGAP003328-PA [Anopheles gambia...  171    2e-41  UG
gi|157134349|ref|XP 001663253.1|  NADH dehydrogenase, putative...  169    5e-41  G
gi|170046809|ref|XP 001850941.1|  NADH dehydrogenase [Culex pi...  165    9e-40  G
gi|19922002|ref|NP 610629.1|  CG7712 CG7712-PA [Drosophila mel...  160    2e-38  UG
gi|125808965|ref|XP 001360938.1|  GA20535-PA [Drosophila pseud...  154    1e-36  G
gi|170041213|ref|XP 001848366.1|  NADH dehydrogenase 1 alpha s...  122    9e-27  G
gi|91079452|ref|XP 969319.1|  PREDICTED: similar to CG7712-PA ...  120    4e-26  UG
gi|156553857|ref|XP 001600564.1|  PREDICTED: similar to NADH d...  116    4e-25  G
gi|90820014|gb|ABD98764.1|  putative NADH-ubiquinone oxidoredu...  112    5e-24
gi|33521688|gb|AAQ21387.1|  NADH-ubiquinone oxidoreductase [Ix...  110    4e-23
gi|66513180|ref|XP 623441.1|  PREDICTED: similar to CG7712-PA ...  101    2e-20  UG
gi|156358613|ref|XP 001624611.1|  predicted protein [Nematoste...  91.3   2e-17  G
gi|41055750|ref|NP 957262.1|  NADH dehydrogenase (ubiquinone) ...  89.4   7e-17  UG
gi|47217026|emb|CAG01654.1|  unnamed protein product [Tetraodo...  88.6   1e-16
gi|51317370|ref|NP 002481.2|  NADH dehydrogenase (ubiquinone) ...  85.1   1e-15  UG
gi|60652655|gb|AAX29022.1|  NADH dehydrogenase 1 alpha subcomp...  84.7   2e-15
gi|48145545|emb|CAG32995.1|  NDUFA6 [Homo sapiens]                 84.7   2e-15  G
gi|115392053|ref|NP 001065259.1|  NADH dehydrogenase (ubiquino...  84.7   2e-15  UG
gi|60833616|gb|AAX37056.1|  NADH dehydrogenase 1 alpha subcomp...  84.7   2e-15
gi|27663138|ref|XP 235518.1|  PREDICTED: similar to NADH dehyd...  84.7   2e-15  G
gi|115502287|sp|Q0MQA3|NDUA6 PONPY  NADH dehydrogenase [ubiqui...  84.3   2e-15
gi|109094394|ref|XP 001106675.1|  PREDICTED: similar to NADH d...  84.0   3e-15  UG
gi|126339065|ref|XP 001371452.1|  PREDICTED: similar to NADH d...  84.0   3e-15  UG
gi|28461207|ref|NP 786985.1|  NADH dehydrogenase (ubiquinone) ...  83.6   3e-15  UG
gi|148232387|ref|NP 001088970.1|  hypothetical protein LOC4963...  83.6   4e-15  UG
6|ref|XP 001516880.1|  PREDICTED: hypothetical prot...  83.2   4e-15  UG
2|ref|XP 001746412.1|  predicted protein [Monosiga ...  83.2   5e-15  G
0|ref|XP 001500539.1|  PREDICTED: similar to NDUFA6...  83.2   5e-15  UG
gi|60813144|gb|AAX36248.1|  NADH dehydrogenase 1 alpha subcomp...  82.8   6e-15
gi|60825365|gb|AAX36716.1|  NADH dehydrogenase 1 alpha subcomp...  82.8   6e-15
gi|73969393|ref|XP 531712.2|  PREDICTED: similar to NADH dehyd...  82.0   1e-14  UG
...  81.6   1e-14  UG
...  81.3   2e-14
...  81.3   2e-14  UG
gi|118082637|ref|XP 425471.2|  PREDICTED: hypothetical protein...  80.5   3e-14  UG
gi|13385492|ref|NP 080263.1|  NADH dehydrogenase (ubiquinone) ...  79.7   5e-14  UG
861 protein [Schistosoma j...  72.4   9e-12
edicted protein [Laccaria ...  61.6   2e-08  G
d protein product [Podospo...  60.8   3e-08
d protein product [Vitis v...  57.4   3e-07
gi|169854690|ref|XP 001834019.1|  hypothetical protein CC1G_09...  55.1   1e-06  G
gi|71013394|ref|XP 758584.1|  hypothetical protein UM02437.1 [...  54.7   2e-06  G
```

BLASTP results: a hit?

The second protein in the output is a “conceptual translation” (86% identical over 98 aa):



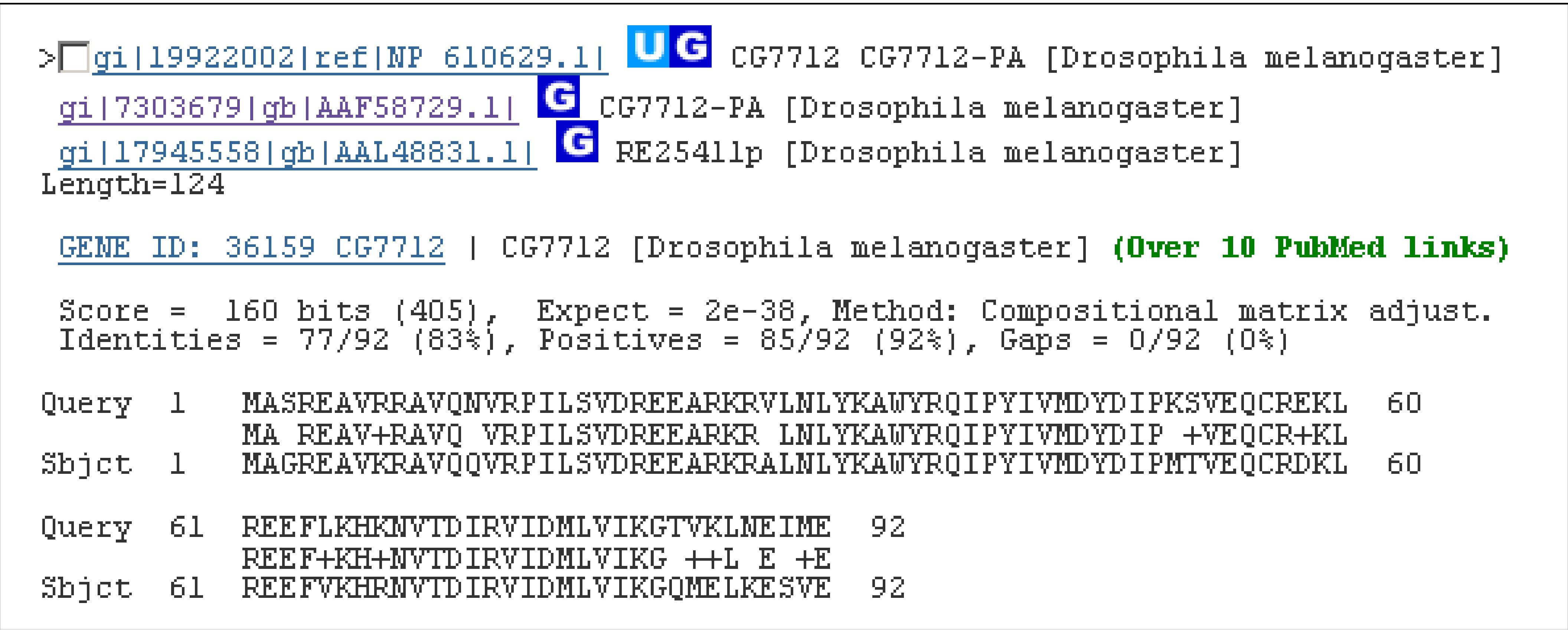
□ 1: XP 306101. Reports AGAP012533-PA [An...[gi:158284321]

```
LOCUS       XP_306101                128 aa            linear   INV 16-OCT-2007
DEFINITION  AGAP012533-PA [Anopheles gambiae str. PEST].
ACCESSION   XP_306101
VERSION     XP_306101.3  GI:158284321
DBSOURCE    REFSEQ: accession XM 306101.3
KEYWORDS    .
SOURCE      Anopheles gambiae str. PEST
  ORGANISM  Anopheles gambiae str. PEST
            Eukaryota; Metazoa; Arthropoda; Hexapoda; Insecta; Pterygota;
            Neoptera; Endopterygota; Diptera; Nematocera; Culicoidea;
            Culicidae; Anophelinae; Anopheles.
REFERENCE   1  (residues 1 to 128)
  AUTHORS   Hammond,M.
  CONSRTM   The Anopheles Genome Sequencing Consortium
  TITLE     The genome sequence of the malaria mosquito Anopheles gambiae
  JOURNAL   Unpublished
COMMENT     PROVISIONAL REFSEQ: This record has not yet been subject to final
            NCBI review. The reference sequence was derived from EAA02579.
            On Oct 15, 2007 this sequence version replaced gi:58374760.
            COMPLETENESS: incomplete on the amino end.
            Method: conceptual translation.
FEATURES             Location/Qualifiers
     source          1..128
                     /organism="Anopheles gambiae str. PEST"
```

NOT USEFUL!!

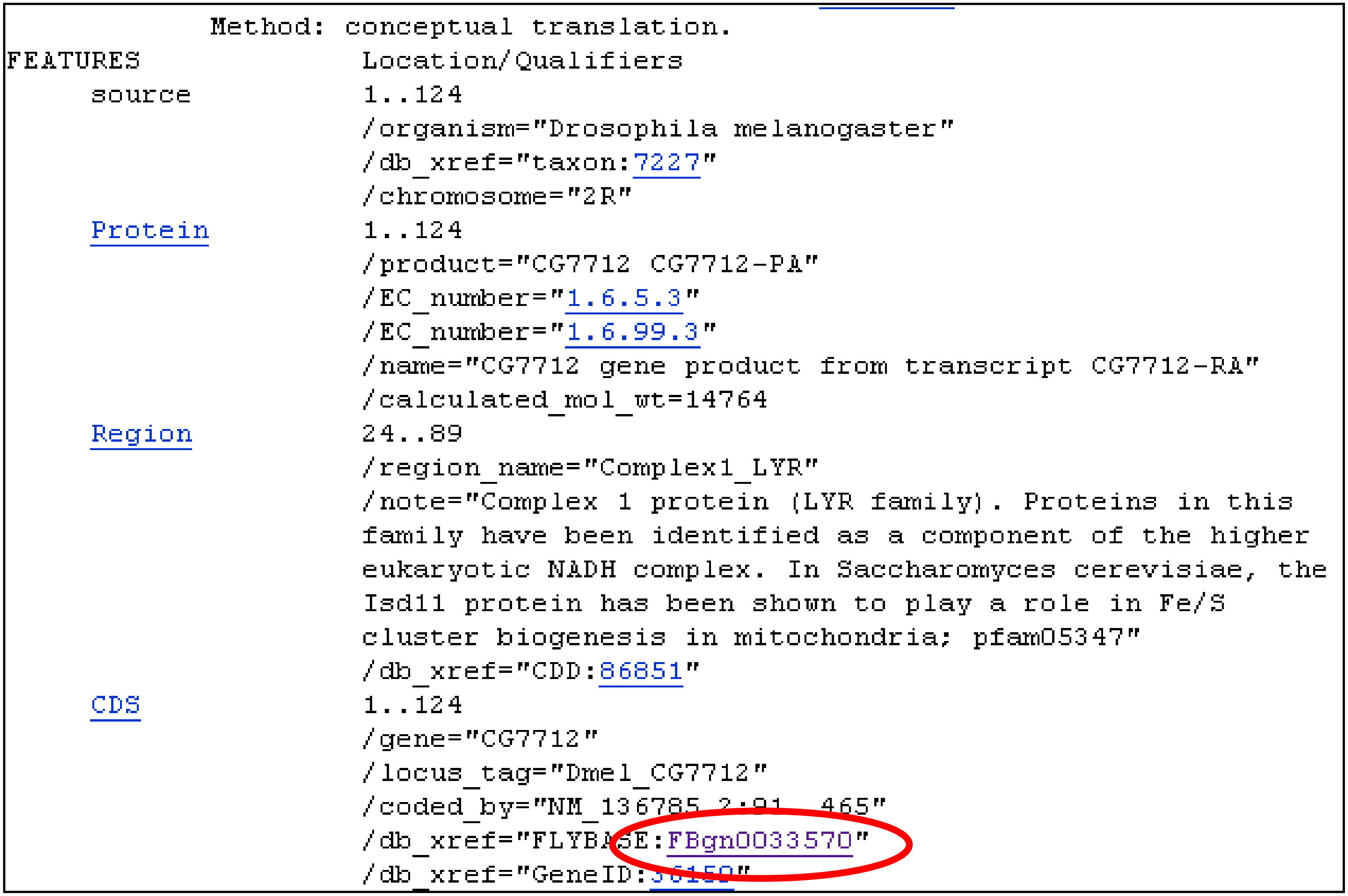
Characterized match?

- The first hit to an annotated protein is to #6 in the list, a Drosophila sequence:



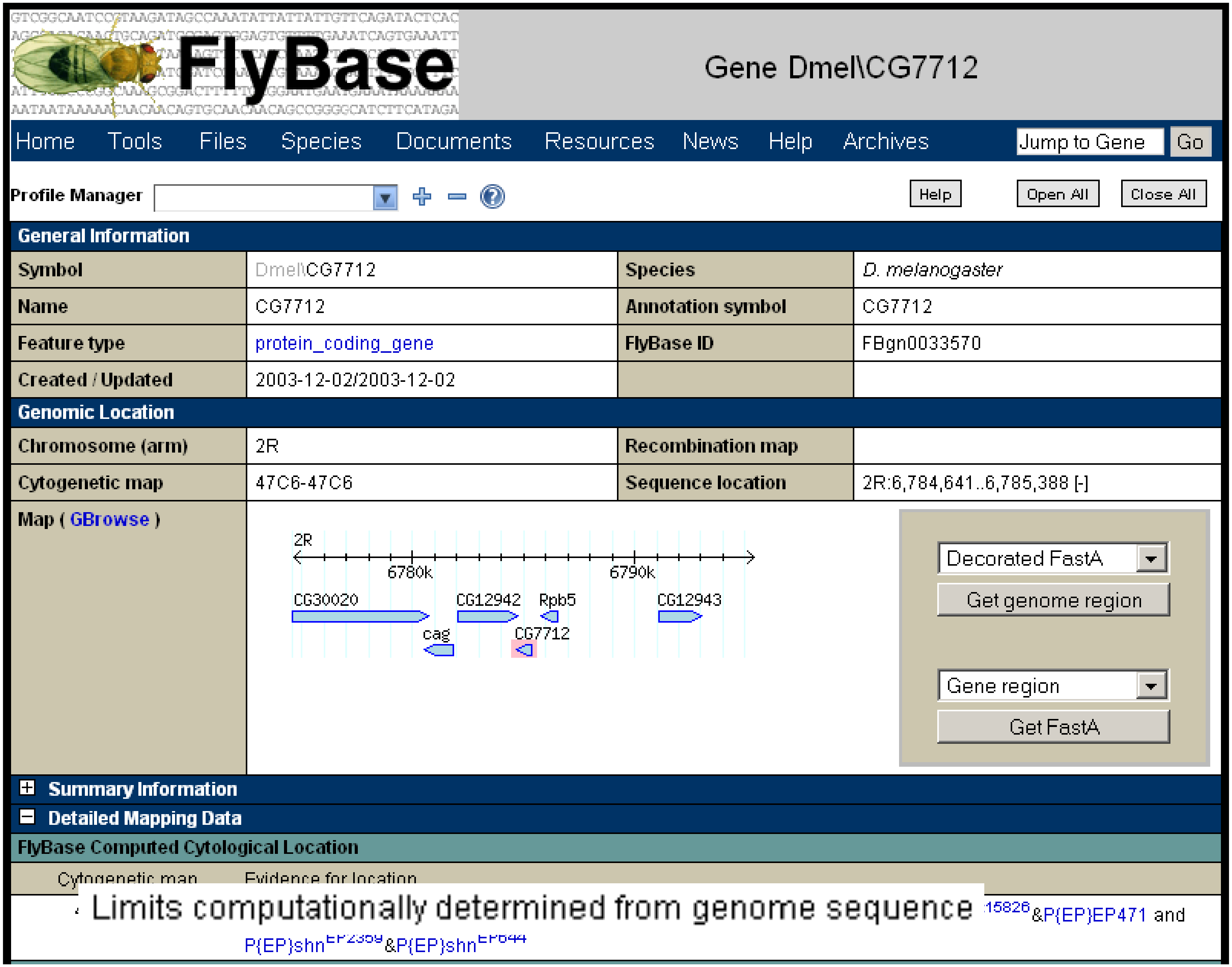
```
>□gi|19922002|ref|NP_610629.1| UG CG7712 CG7712-PA [Drosophila melanogaster]
 gi|7303679|gb|AAF58729.1| G CG7712-PA [Drosophila melanogaster]
 gi|17945558|gb|AAL48831.1| G RE25411p [Drosophila melanogaster]
Length=124

 GENE ID: 36159 CG7712 | CG7712 [Drosophila melanogaster] (Over 10 PubMed links)

 Score =  160 bits (405),  Expect = 2e-38, Method: Compositional matrix adjust.
 Identities = 77/92 (83%), Positives = 85/92 (92%), Gaps = 0/92 (0%)

Query  1   MASREAVRRAVQNVRPILSVDREEARKRVLNLYKAWYRQIPYIVMDYDIPKSVEQCREKL  60
           MA REAV+RAVQ VRPILSVDREEARKR LNLYKAWYRQIPYIVMDYDIP +VEQCR+KL
Sbjct  1   MAGREAVKRAVQQVRPILSVDREEARKRALNLYKAWYRQIPYIVMDYDIPMTVEQCRDKL  60

Query  61  REEFLKHKNVTDIRVIDMLVIKGTVKLNEIME  92
           REEF+KH+NVTDIRVIDMLVIKG ++L E +E
Sbjct  61  REEFVKHRNVTDIRVIDMLVIKGQMELKESVE  92
```

Evidence for validity of the protein it matches?



```
            Method: conceptual translation.
FEATURES             Location/Qualifiers
     source          1..124
                     /organism="Drosophila melanogaster"
                     /db_xref="taxon:7227"
                     /chromosome="2R"
     Protein         1..124
                     /product="CG7712 CG7712-PA"
                     /EC_number="1.6.5.3"
                     /EC_number="1.6.99.3"
                     /name="CG7712 gene product from transcript CG7712-RA"
                     /calculated_mol_wt=14764
     Region          24..89
                     /region_name="Complex1_LYR"
                     /note="Complex 1 protein (LYR family). Proteins in this
                     family have been identified as a component of the higher
                     eukaryotic NADH complex. In Saccharomyces cerevisiae, the
                     Isd11 protein has been shown to play a role in Fe/S
                     cluster biogenesis in mitochondria; pfam05347"
                     /db_xref="CDD:86851"
     CDS             1..124
                     /gene="CG7712"
                     /locus_tag="Dmel_CG7712"
                     /coded_by="NM_136785.2:91..465"
                     /db_xref="FLYBASE:FBgn0033570"
                     /db_xref="GeneID:36150"
```

Exploring the match



Gene Dmel\CG7712

General Information

Symbol	Dmel\CG7712	**Species**	*D. melanogaster*
Name	CG7712	**Annotation symbol**	CG7712
Feature type	protein_coding_gene	**FlyBase ID**	FBgn0033570
Created / Updated	2003-12-02/2003-12-02		

Genomic Location

Chromosome (arm)	2R	**Recombination map**	
Cytogenetic map	47C6-47C6	**Sequence location**	2R:6,784,641..6,785,388 [-]
Map (GBrowse)			

Summary Information

Detailed Mapping Data

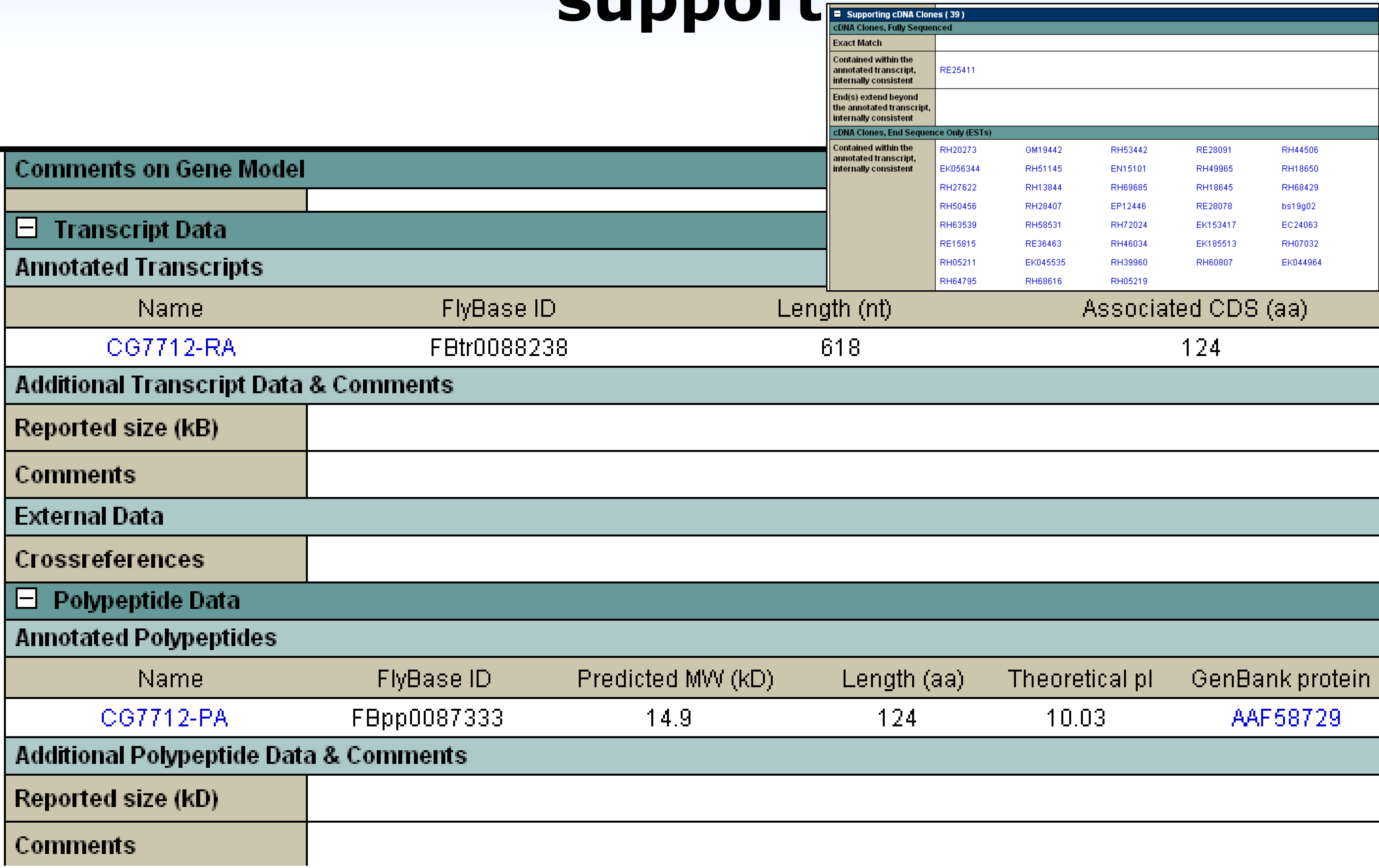
FlyBase Computed Cytological Location

Cytogenetic map Evidence for location

Limits computationally determined from genome sequence

15826 &P{EP}EP471 and P{EP}shn^{EP2359}&P{EP}shn^{EP644}

Drosophila match has transcript support



Comments on Gene Model

⊟ Transcript Data

Annotated Transcripts

Name	FlyBase ID	Length (nt)	Associated CDS (aa)
CG7712-RA	FBtr0088238	618	124

Additional Transcript Data & Comments

Reported size (kB)	
Comments	

External Data

Crossreferences	

⊟ Polypeptide Data

Annotated Polypeptides

Name	FlyBase ID	Predicted MW (kD)	Length (aa)	Theoretical pI	GenBank protein
CG7712-PA	FBpp0087333	14.9	124	10.03	AAF58729

Additional Polypeptide Data & Comments

Reported size (kD)	
Comments	

⊟ Supporting cDNA Clones (39)

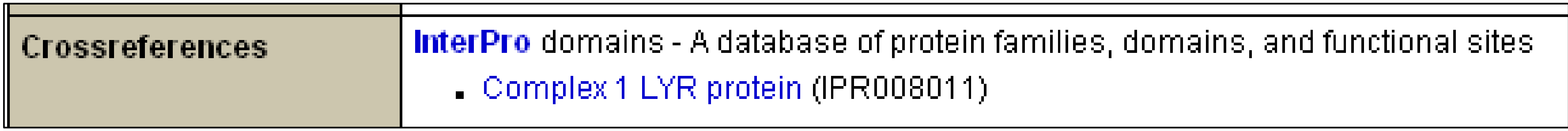
cDNA Clones, Fully Sequenced

Exact Match	
Contained within the annotated transcript, internally consistent	RE25411
End(s) extend beyond the annotated transcript, internally consistent	

cDNA Clones, End Sequence Only (ESTs)

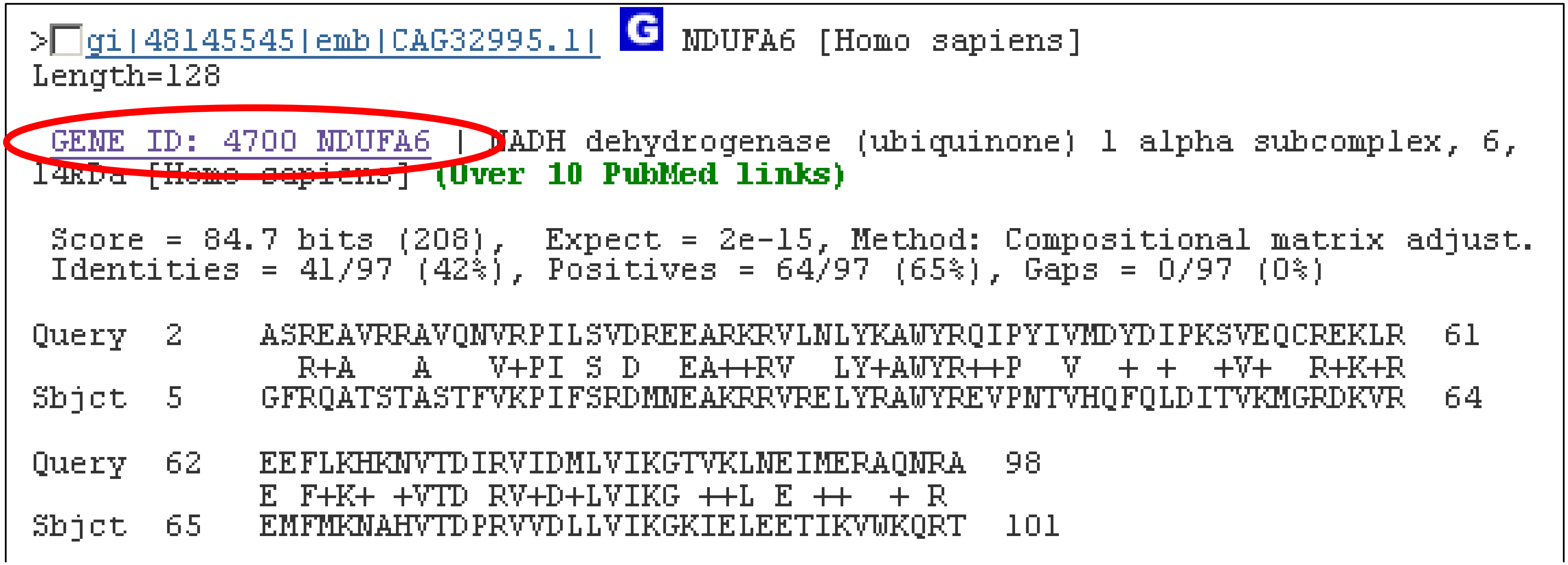
Contained within the annotated transcript, internally consistent	RH20273	GM19442	RH53442	RE28091	RH44506
	EK056344	RH51145	EN15101	RH49965	RH18650
	RH27622	RH13844	RH69685	RH18645	RH68429
	RH50456	RH28407	EP12446	RE28078	bs19g02
	RH63539	RH58531	RH72024	EK153417	EC24063
	RE15815	RE36463	RH46034	EK185513	RH07032
	RH05211	EK045535	RH39960	RH60807	EK044964
	RH64795	RH68616	RH05219		

Match is an expressed protein, with LYR domain



Crossreferences	InterPro domains - A database of protein families, domains, and functional sites ▪ Complex 1 LYR protein (IPR008011)

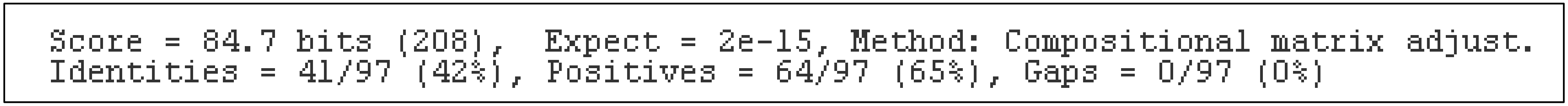
After all of that investigation, we have to conclude that this is not a “characterized match.” We continue down the BLASTP output to #19:



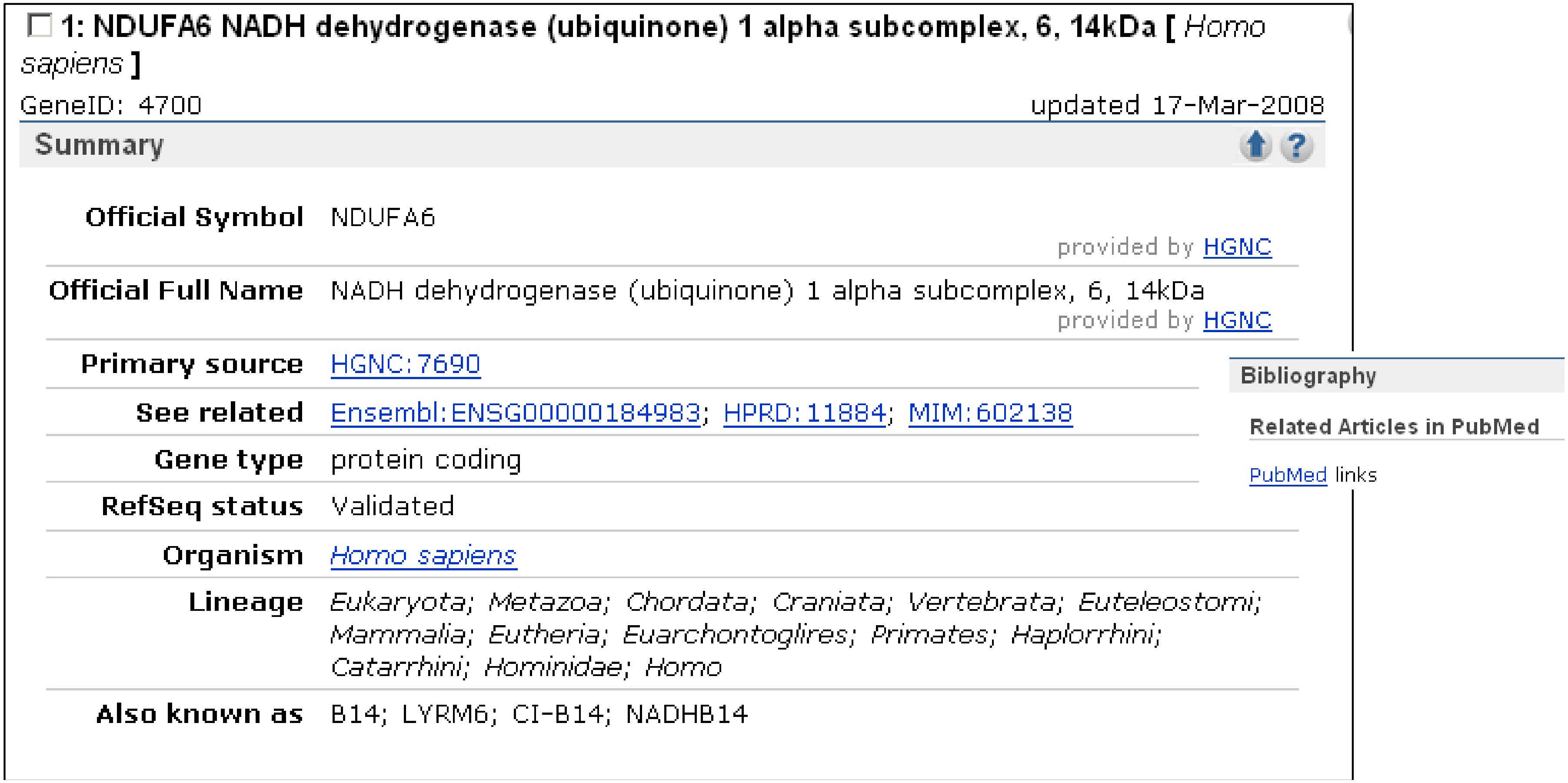
```
>□gi|48145545|emb|CAG32995.1|  G  NDUFA6 [Homo sapiens]
Length=128

 GENE ID: 4700 NDUFA6 | NADH dehydrogenase (ubiquinone) 1 alpha subcomplex, 6,
14kDa [Homo sapiens] (Over 10 PubMed links)

 Score = 84.7 bits (208),  Expect = 2e-15, Method: Compositional matrix adjust.
 Identities = 41/97 (42%), Positives = 64/97 (65%), Gaps = 0/97 (0%)

Query  2    ASREAVRRAVQNVRPILSVDREEARKRVLNLYKAWYRQIPYIVMDYDIPKSVEQCREKLR  61
              R+A   A  V+PI S D  EA++RV  LY+AWYR++P  V  + +  +V+  R+K+R
Sbjct  5    GFRQATSTASTFVKPIFSRDMNEAKRRVRELYRAWYREVPNTVHQFQLDITVKMGRDKVR  64

Query  62   EEFLKHKNVTDIRVIDMLVIKGTVKLNEIMERAQNRA  98
            E F+K+ +VTD RV+D+LVIKG ++L E ++  + R
Sbjct  65   EMFMKNAHVTDPRVVDLLVIKGKIELEETIKVWKQRT  101
```

Investigating this match

This match is at 41% identity over 76% of the length of the matching protein sequence:



```
Score = 84.7 bits (208),  Expect = 2e-15, Method: Compositional matrix adjust.
Identities = 41/97 (42%), Positives = 64/97 (65%), Gaps = 0/97 (0%)
```



☐ **1: NDUFA6 NADH dehydrogenase (ubiquinone) 1 alpha subcomplex, 6, 14kDa [** *Homo sapiens* **]**

GeneID: 4700 updated 17-Mar-2008

Summary ⬆ ?

Official Symbol	NDUFA6 provided by HGNC
Official Full Name	NADH dehydrogenase (ubiquinone) 1 alpha subcomplex, 6, 14kDa provided by HGNC
Primary source	HGNC:7690
See related	Ensembl:ENSG00000184983; HPRD:11884; MIM:602138
Gene type	protein coding
RefSeq status	Validated
Organism	*Homo sapiens*
Lineage	*Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini; Catarrhini; Hominidae; Homo*
Also known as	B14; LYRM6; CI-B14; NADHB14

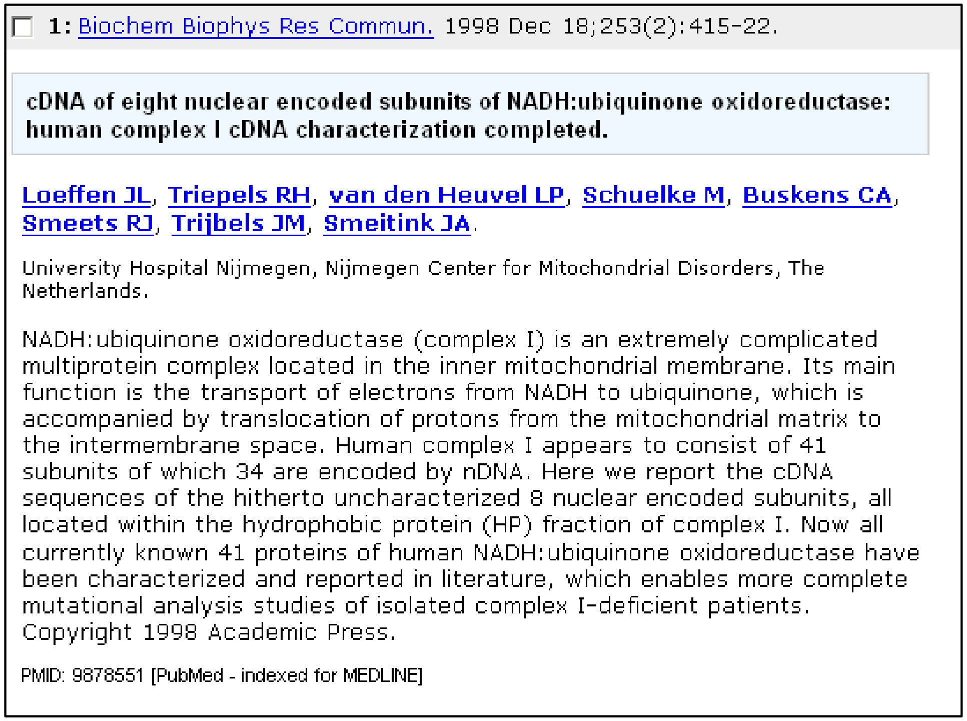
Bibliography

Related Articles in PubMed

PubMed links

Looking at the literature

We now verify that this protein has been characterized, and constitutes a valid characterized match.



□ **1:** Biochem Biophys Res Commun. 1998 Dec 18;253(2):415-22.

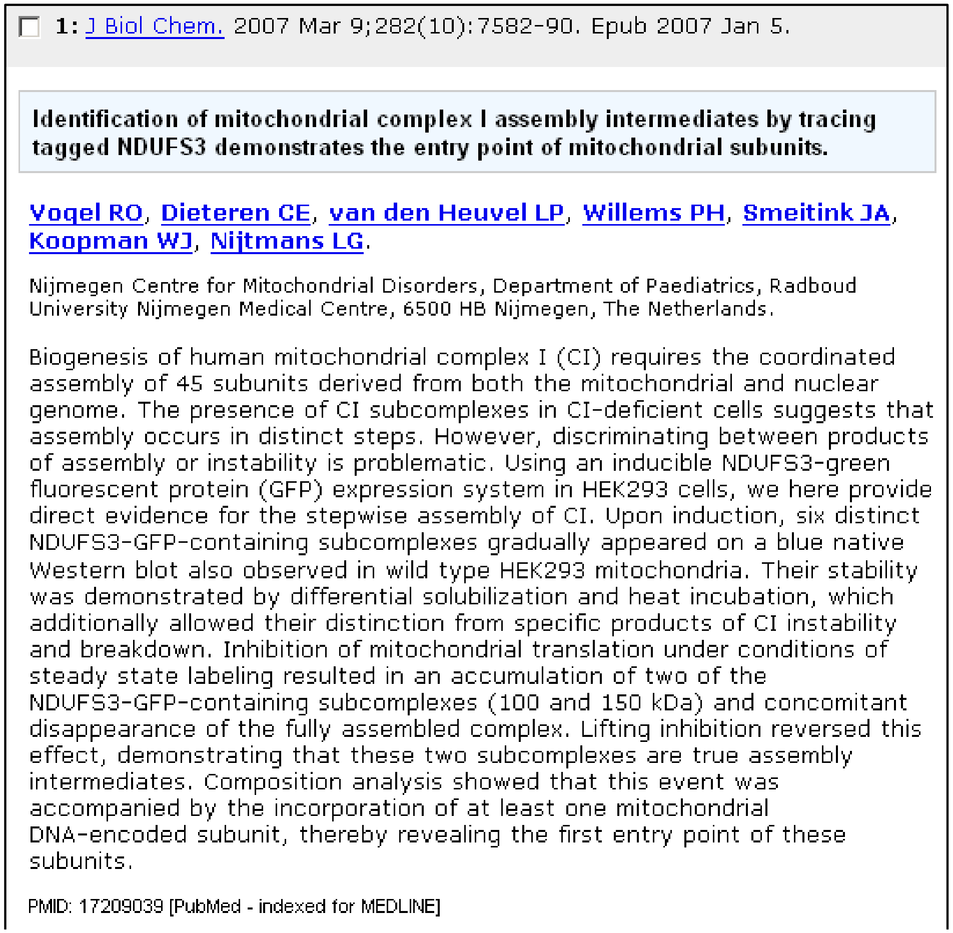
cDNA of eight nuclear encoded subunits of NADH:ubiquinone oxidoreductase: human complex I cDNA characterization completed.

Loeffen JL, Triepels RH, van den Heuvel LP, Schuelke M, Buskens CA, Smeets RJ, Trijbels JM, Smeitink JA.

University Hospital Nijmegen, Nijmegen Center for Mitochondrial Disorders, The Netherlands.

NADH:ubiquinone oxidoreductase (complex I) is an extremely complicated multiprotein complex located in the inner mitochondrial membrane. Its main function is the transport of electrons from NADH to ubiquinone, which is accompanied by translocation of protons from the mitochondrial matrix to the intermembrane space. Human complex I appears to consist of 41 subunits of which 34 are encoded by nDNA. Here we report the cDNA sequences of the hitherto uncharacterized 8 nuclear encoded subunits, all located within the hydrophobic protein (HP) fraction of complex I. Now all currently known 41 proteins of human NADH:ubiquinone oxidoreductase have been characterized and reported in literature, which enables more complete mutational analysis studies of isolated complex I-deficient patients. Copyright 1998 Academic Press.

PMID: 9878551 [PubMed - indexed for MEDLINE]



□ **1:** J Biol Chem. 2007 Mar 9;282(10):7582-90. Epub 2007 Jan 5.

Identification of mitochondrial complex I assembly intermediates by tracing tagged NDUFS3 demonstrates the entry point of mitochondrial subunits.

Vogel RO, Dieteren CE, van den Heuvel LP, Willems PH, Smeitink JA, Koopman WJ, Nijtmans LG.

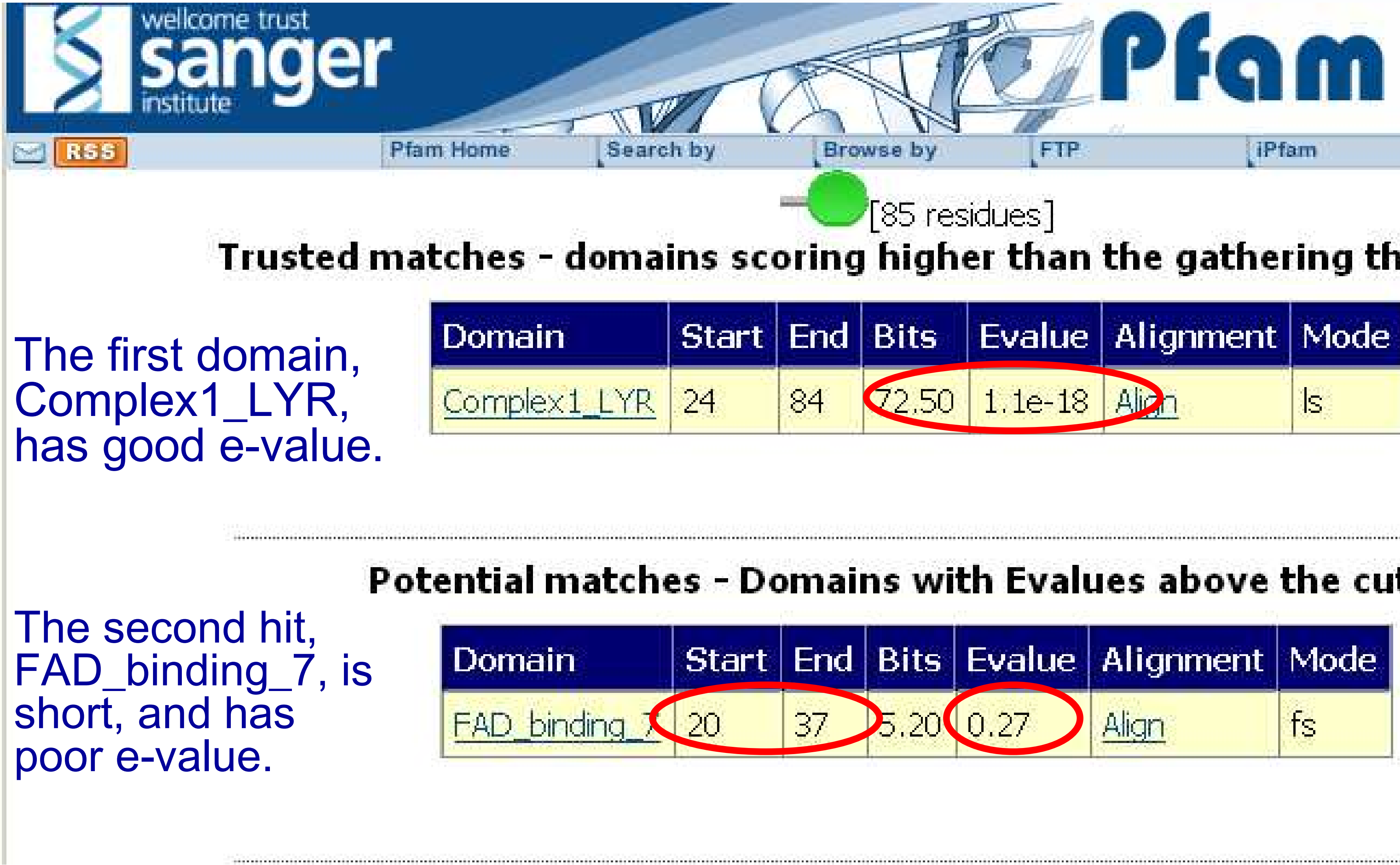
Nijmegen Centre for Mitochondrial Disorders, Department of Paediatrics, Radboud University Nijmegen Medical Centre, 6500 HB Nijmegen, The Netherlands.

Biogenesis of human mitochondrial complex I (CI) requires the coordinated assembly of 45 subunits derived from both the mitochondrial and nuclear genome. The presence of CI subcomplexes in CI-deficient cells suggests that assembly occurs in distinct steps. However, discriminating between products of assembly or instability is problematic. Using an inducible NDUFS3-green fluorescent protein (GFP) expression system in HEK293 cells, we here provide direct evidence for the stepwise assembly of CI. Upon induction, six distinct NDUFS3-GFP-containing subcomplexes gradually appeared on a blue native Western blot also observed in wild type HEK293 mitochondria. Their stability was demonstrated by differential solubilization and heat incubation, which additionally allowed their distinction from specific products of CI instability and breakdown. Inhibition of mitochondrial translation under conditions of steady state labeling resulted in an accumulation of two of the NDUFS3-GFP-containing subcomplexes (100 and 150 kDa) and concomitant disappearance of the fully assembled complex. Lifting inhibition reversed this effect, demonstrating that these two subcomplexes are true assembly intermediates. Composition analysis showed that this event was accompanied by the incorporation of at least one mitochondrial DNA-encoded subunit, thereby revealing the first entry point of these subunits.

PMID: 17209039 [PubMed - indexed for MEDLINE]

Example 2: HMM search on our sequence

http://www.sanger.ac.uk/Software/Pfam/



[85 residues]

Trusted matches - domains scoring higher than the gathering threshold (A

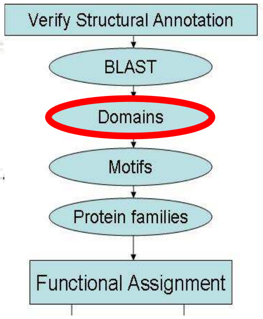
Domain	Start	End	Bits	Evalue	Alignment	Mode
Complex1_LYR	24	84	72.50	1.1e-18	Align	ls

The first domain, Complex1_LYR, has good e-value.

Potential matches - Domains with Evalues above the cutoff

Domain	Start	End	Bits	Evalue	Alignment	Mode
FAD_binding_7	20	37	5.20	0.27	Align	fs

The second hit, FAD_binding_7, is short, and has poor e-value.

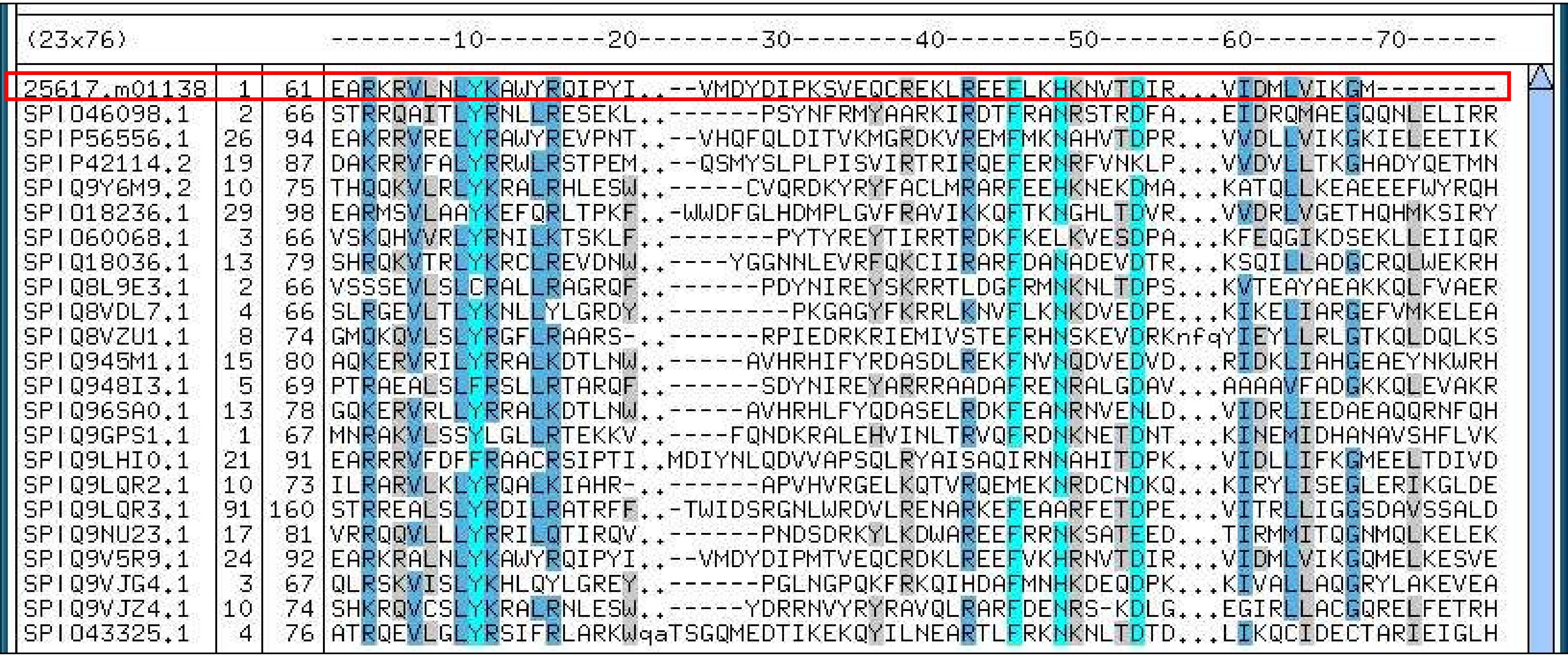


Examine HMM evidence for our sequence (Belvu tool to display alignment)

http://sonnhammer.sbc.su.se/Belvu.html

View the HMM Alignment:

The “seed” is the set of sequences that are used to make up the statistical model of the domain (HMM). Examine our sequence aligned to the SEED (at the Pfam site).

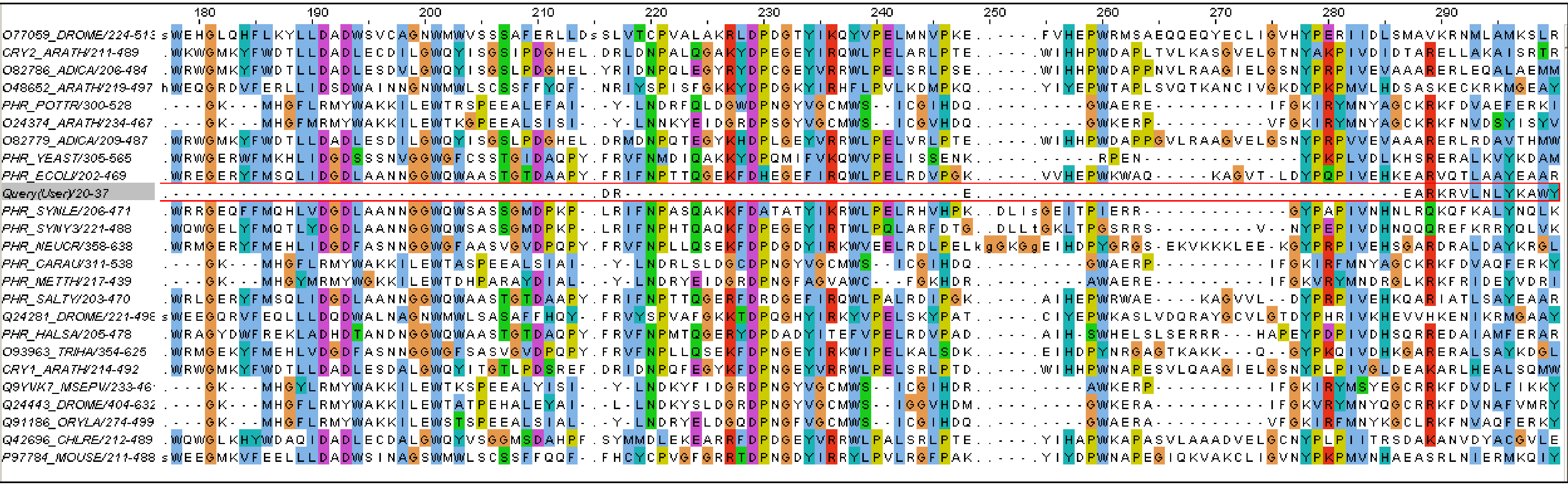


```
(23x76)                    --------10--------20--------30--------40--------50--------60--------70------
25617.m01138    1   61 EARKRVLNLYKAWYRQIPYI..--VMDYDIPKSVEQCREKLREEFLKHKNVTDIR...VIDMLVIKGM--------
SP|O46098.1     2   66 STRRQAITLYRNLLRESEKL..------PSYNFRMYAARKIRDTFRANRSTRDFA...EIDRQMAEGQQNLELIRR
SP|P56556.1    26   94 EAKRRVRELYRAWYREVPNT..--VHQFQLDITVKMGRDKVREMFMKNAHVTDPR...VVDLLVIKGKIELEETIK
SP|P42114.2    19   87 DAKRRVFALYRRWLRSTPEM..--QSMYSLPLPISVIRTRIRQEFERNRFVNKLP...VVDVLLTKGHADYQETMN
SP|Q9Y6M9.2    10   75 THQQKVLRLYKRALRHLESW..-----CVQRDKYRYFACLMRARFEEHKNEKDMA...KATQLLKEAEEEFWYRQH
SP|O18236.1    29   98 EARMSVLAAYKEFQRLTPKF..-WWDFGLHDMPLGVFRAVIKKQFTKNGHLTDVR...VVDRLVGETHQHMKSIRY
SP|O60068.1     3   66 VSKQHVVRLYRNILKTSKLF..-------PYTYREYTIRRTRDKFKELKVESDPA...KFEQGIKDSEKLLEIIQR
SP|Q18036.1    13   79 SHRQKVTRLYKRCLREVDNW..----YGGNNLEVRFQKCIIRARFDANADEVDTR...KSQILLADGCRQLWEKRH
SP|Q8L9E3.1     2   66 VSSSEVLSLCRALLRAGRQF..------PDYNIREYSKRRTLDGFRMNKNLTDPS...KVTEAYAEAKKQLFVAER
SP|Q8VDL7.1     4   66 SLRGEVLTLYKNLLYLGRDY..--------PKGAGYFKRRLKNVFLKNKDVEDPE...KIKELIARGEFVMKELEA
SP|Q8VZU1.1     8   74 GMQKQVLSLYRGFLRAARS-..------RPIEDRKRIEMIVSTEFRHNSKEVDRKnfqYIEYLLRLGTKQLDQLKS
SP|Q945M1.1    15   80 AQKERVRILYRRALKDTLNW..-----AVHRHIFYRDASDLREKFNVNQDVEDVD...RIDKLIAHGEAEYNKWRH
SP|Q948I3.1     5   69 PTRAEALSLFRSLLRTARQF..------SDYNIREYARRRAADAFRENRALGDAV...AAAAVFADGKKQLEVAKR
SP|Q96SA0.1    13   78 GQKERVRLLYRRALKDTLNW..-----AVHRHLFYQDASELRDKFEANRNVENLD...VIDRLIEDAEAQQRNFQH
SP|Q9GPS1.1     1   67 MNRAKVLSSYLGLLRTEKKV..----FQNDKRALEHVINLTRVQFRDNKNETDNT...KINEMIDHANAVSHFLVK
SP|Q9LHI0.1    21   91 EARRRVFDFFRAACRSIPTI..MDIYNLQDVVAPSQLRYAISAQIRNNAHITDPK...VIDLLIFKGMEELTDIVD
SP|Q9LQR2.1    10   73 ILRARVLKLYRQALKIAHR-..------APVHVRGELKQTVRQEMEKNRDCNDKQ...KIRYLISEGLERIKGLDE
SP|Q9LQR3.1    91  160 STRREALSLYRDILRATRFF..-TWIDSRGNLWRDVLRENARKEFEAARFETDPE...VITRLLIGGSDAVSSALD
SP|Q9NU23.1    17   81 VRRQQVLLLYRRILQTIRQV..------PNDSDRKYLKDWAREEFRRNKSATEED...TIRMMITQGNMQLKELEK
SP|Q9V5R9.1    24   92 EARKRALNLYKAWYRQIPYI..--VMDYDIPMTVEQCRDKLREEFVKHRNVTDIR...VIDMLVIKGQMELKESVE
SP|Q9VJG4.1     3   67 QLRSKVISLYKHLQYLGREY..------PGLNGPQKFRKQIHDAFMNHKDEQDPK...KIVALLAQGRYLAKEVEA
SP|Q9VJZ4.1    10   74 SHKRQVCSLYKRALRNLESW..-----YDRRNVYRYRAVQLRARFDENRS-KDLG...EGIRLLACGQRELFETRH
SP|O43325.1     4   76 ATRQEVLGLYRSIFRLARKWqaTSGQMEDTIKEKQYILNEARTLFRKNKNLTDTD...LIKQCIDECTARIEIGLH
```

Pfam: Jalview tool to verify alignment to seed

http://www.jalview.org/

The second domain alignment shows us why the score is low. Much of the sequence of the domain is missing!



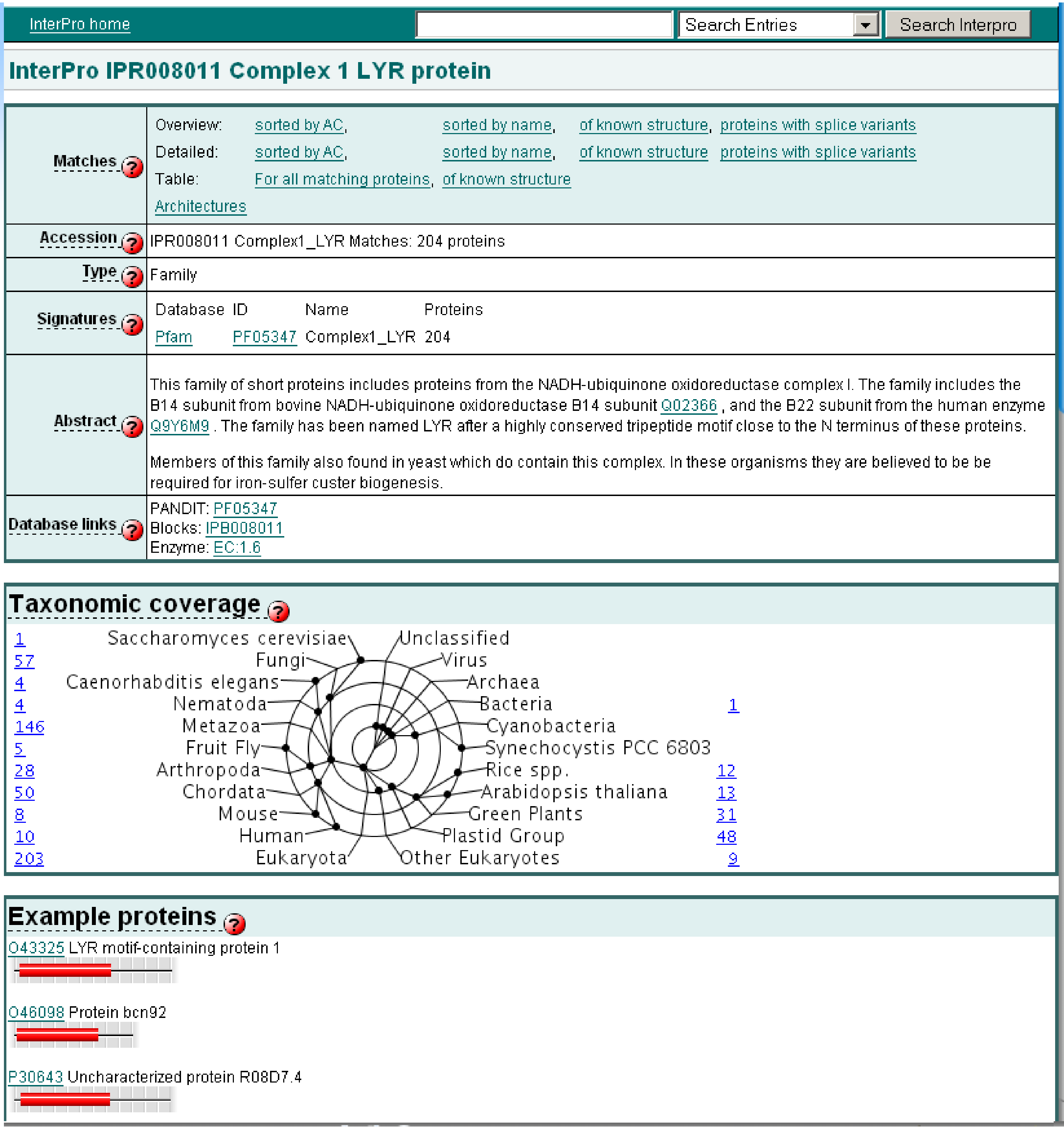
```
                        180       190       200       210       220       230       240       250       260       270       280       290
O77059_DROME/224-513 sWEHGLQHFLKYLLDADWSVCAGNWMWVSSSAFERLLDsSLVTCPVALAKRLDPDGTYIKQYVPELMNVPKE.----.FVHEPWRMSAEQQEQYECLIGVHYPERIIDLSMAVKRNMLAMKSLR
CRY2_ARATH/211-489   .WKWGMKYFWDTLLDADLECDILGWQYISGSIPDGHEL.DRLDNPALQGAKYDPEGEYIRQWLPELARLPTE.----.WIHHPWDAPLTVLKASGVELGTNYAKPIVDIDTARELLAKAISRTR
O82786_ADICA/206-484 .WRWGMKYFWDTLLDADLESDVLGWQYISGSLPDGHEL.YRIDNPQLEGYRYDPCGEYVRRWLPELSRLPSE.----.WIHHPWDAPPNVLRAAGIELGSNYPRPIVEVAAARERLEQALAEMM
O48652_ARATH/219-497 hWEQGRDVFERLLIDSDWAINNGNWMWLSCSSFFYQF-.NRIYSPISFGKKYDPDGKYIRHFLPVLKDMPKQ.----.YIYEPWTAPLSVQTKANCIVGKDYPKPMVLHDSASKECKRKMGEAY
PHR_POTTR/300-528    .----GK---MHGFLRMYWAKKILEWTRSPEEALEFAI-.-Y-LNDRFQLDGWDPNGYVGCMWS--ICGIHDQ.----.----GWAERE----------IFGKIRYMNYAGCKRKFDVAEFERKI
O24374_ARATH/234-467 .----GK---MHGFMRMYWAKKILEWTKGPEEALSISI-.-Y-LNNKYEIDGRDPSGYVGCMWS--ICGVHDQ.----.----GWKERP----------VFGKIRYMNYAGCKRKFNVDSYISYV
O82779_ADICA/209-487 .WRWGMKYFWDTLLDADLESDILGWQYISGSLPDGHEL.DRMDNPQTEGYKHDPLGEYVRRWLPELVRLPTE.----.WIHHPWDAPPGVLRAAGVELGSNYPRPVVEVAAARERLRDAVTHMW
PHR_YEAST/305-565    .WRWGERWFMKHLIDGDSSSNVGGWGFCSSTGIDAQPY.FRVFNMDIQAKKYDPQMIFVKQWVPELISSENK.----.-----RPEN--------------YPKPLVDLKHSRERALKVYKDAM
PHR_ECOLI/202-469    .WREGERYFMSQLIDGDLAANNGGWQWAASTGTDAAPY.FRIFNPTTQGEKFDHEGEFIRQWLPELRDVPGK.----.VVHEPWKWAQ-----KAGVT-LDYPQPIVEHKEARVQTLAAYEAAR
Query(User)/20-37    .-------------------------------------.DR------------------------------E.----.--------------------------------EARKRVLNLYKAWY
PHR_SYNLE/206-471    .WRRGEQFFMQHLVDGDLAANNGGWQWSASSGMDPKP-.LRIFNPASQAKKFDATATYIKRWLPELRHVHPK..DLIsGEITPIERR-------------GYPAPIVNHNLRQKQFKALYNQLK
PHR_SYNY3/221-488    .WQWGELYFMQTLYDGDLAANNGGWQWSASSGMDPKP-.LRIFNPHTQAQKFDPEGEYIRTWLPQLARFDTG..DLLtGKLTPGSRRS---------V--NYPEPIVDHNQQQREFKRRYQLVK
PHR_NEUCR/358-638    .WRMGERYFMEHLIDGDFASNNGGWGFAASVGVDPQPY.FRVFNPLLQSEKFDPDGDYIRKWVEELRDLPELkgGKGgEIHDPYGRGS-EKVKKKLEE-KGYPRPIVEHSGARDRALDAYKRGL
PHR_CARAU/311-538    .----GK---MHGFLRMYWAKKILEWTASPEEALSIAI-.-Y-LNDRLSLDGCDPNGYVGCMWS--ICGIHDQ.----.----GWAERP----------IFGKIRFMNYAGCKRKFDVAQFERKY
PHR_METTH/217-439    .----GK---MHGYMRMYWGKKILEWTDHPARAYDIAL-.-Y-LNDRYEIDGRDPNGFAGVAWC---FGKHDR.----.----AWAERE----------IFGKVRYMNDRGLKRKFRIDEYVDRI
PHR_SALTY/203-470    .WRLGERYFMSQLIDGDLAANNGGWQWAASTGTDAAPY.FRIFNPTTQGERFDRDGEFIRQWLPALRDIPGK.----.AIHEPWRWAE----KAGVVL--DYPRPIVEHKQARIATLSAYEAAR
Q24281_DROME/221-498 sWEEGQRVFEQLLLDQDWALNAGNWMWLSASAFFHQY-.FRVYSPVAFGKKTDPQGHYIRKYVPELSKYPAT.----.CIYEPWKASLVDQRAYGCVLGTDYPHRIVKHEVVHKENIKRMGAAY
PHR_HALSA/205-478    .WRAGYDWFREKLADHDTANDNGGWQWAASTGTDAQPY.FRVFNPMTQGERYDPDADYITEFVPELRDVPAD.----.AIH-SWHELSLSERRR---HAPEYPDPIVDHSQRREDAIAMFERAR
O93963_TRIHA/354-625 .WRMGEKYFMEHLVDGDFASNNGGWGFSASVGVDPQPY.FRVFNPLLQSEKFDPNGEYIRKWIPELKALSDK.----.EIHDPYNRGAGTKAKK---Q--GYPKQIVDHKGARERALSAYKDGL
CRY1_ARATH/214-492   .WRWGMKYFWDTLLDADLESDALGWQYITGTLPDSREF.DRIDNPQFEGYKFDPNGEYVRRWLPELSRLPTD.----.WIHHPWNAPESVLQAAGIELGSNYPLPIVGLDEAKARLHEALSQMW
Q9YVK7_MSEPV/233-46  .----GK---MHGYLRMYWAKKILEWTKSPEEALYISI-.-Y-LNDKYFIDGRDPNGYVGCMWS--ICGIHDR.----.----AWKERP----------IFGKIRYMSYEGCRRKFDVDLFIKKY
Q24443_DROME/404-632 .----GK---MHGFLRMYWAKKILEWTATPEHALEYAI-.-L-LNDKYSLDGRDPNGYVGCMWS--IGGVHDM.----.----GWKERA----------IFGKVRYMNYQGCRRKFDVNAFVMRY
Q91186_ORYLA/274-499 .----GK---MHGFLRMYWAKKILEWSTSPEEALSIAL-.-Y-LNDRYELDGQDPNGFVGCMWS--ICGIHDQ.----.----GWAERA----------VFGKIRFMNYKGCLRKFNVAQFERKY
Q42696_CHLRE/212-489 .WQWGLKHYWDAQIDADLECDALGWQYVSGGMSDAHPF.SYMMDLEKEARRFDPDGEYVRRWLPALSRLPTE.----.YIHAPWKAPASVLAAADVELGCNYPLPIITRSDAKANVDYACGVLE
P97784_MOUSE/211-488 sWEEGMKVFEELLLDADWSINAGSWMWLSCSSFFQQF-.FHCYCPVGFGRRTDPNGDYIRRYLPVLRGFPAK.----.YIYDPWNAPEGIQKVAKCLIGVNYPKPMVNHAEASRLNIERMKQIY
```

Interpro

http://www.ebi.ac.uk/interpro/

InterPro:

Our protein belongs to this family. It has the domain PF05347.



InterPro IPR008011 Complex 1 LYR protein

Matches	Overview: sorted by AC, sorted by name, of known structure, proteins with splice variants Detailed: sorted by AC, sorted by name, of known structure proteins with splice variants Table: For all matching proteins, of known structure Architectures
Accession	IPR008011 Complex1_LYR Matches: 204 proteins
Type	Family
Signatures	Database ID Name Proteins Pfam PF05347 Complex1_LYR 204
Abstract	This family of short proteins includes proteins from the NADH-ubiquinone oxidoreductase complex I. The family includes the B14 subunit from bovine NADH-ubiquinone oxidoreductase B14 subunit Q02366 , and the B22 subunit from the human enzyme Q9Y6M9 . The family has been named LYR after a highly conserved tripeptide motif close to the N terminus of these proteins. Members of this family also found in yeast which do contain this complex. In these organisms they are believed to be be required for iron-sulfer custer biogenesis.
Database links	PANDIT: PF05347 Blocks: IPB008011 Enzyme: EC:1.6

Taxonomic coverage

1	Saccharomyces cerevisiae	Unclassified	
57	Fungi	Virus	
4	Caenorhabditis elegans	Archaea	
4	Nematoda	Bacteria	1
146	Metazoa	Cyanobacteria	
5	Fruit Fly	Synechocystis PCC 6803	
28	Arthropoda	Rice spp.	12
50	Chordata	Arabidopsis thaliana	13
8	Mouse	Green Plants	31
10	Human	Plastid Group	48
203	Eukaryota	Other Eukaryotes	9

Example proteins

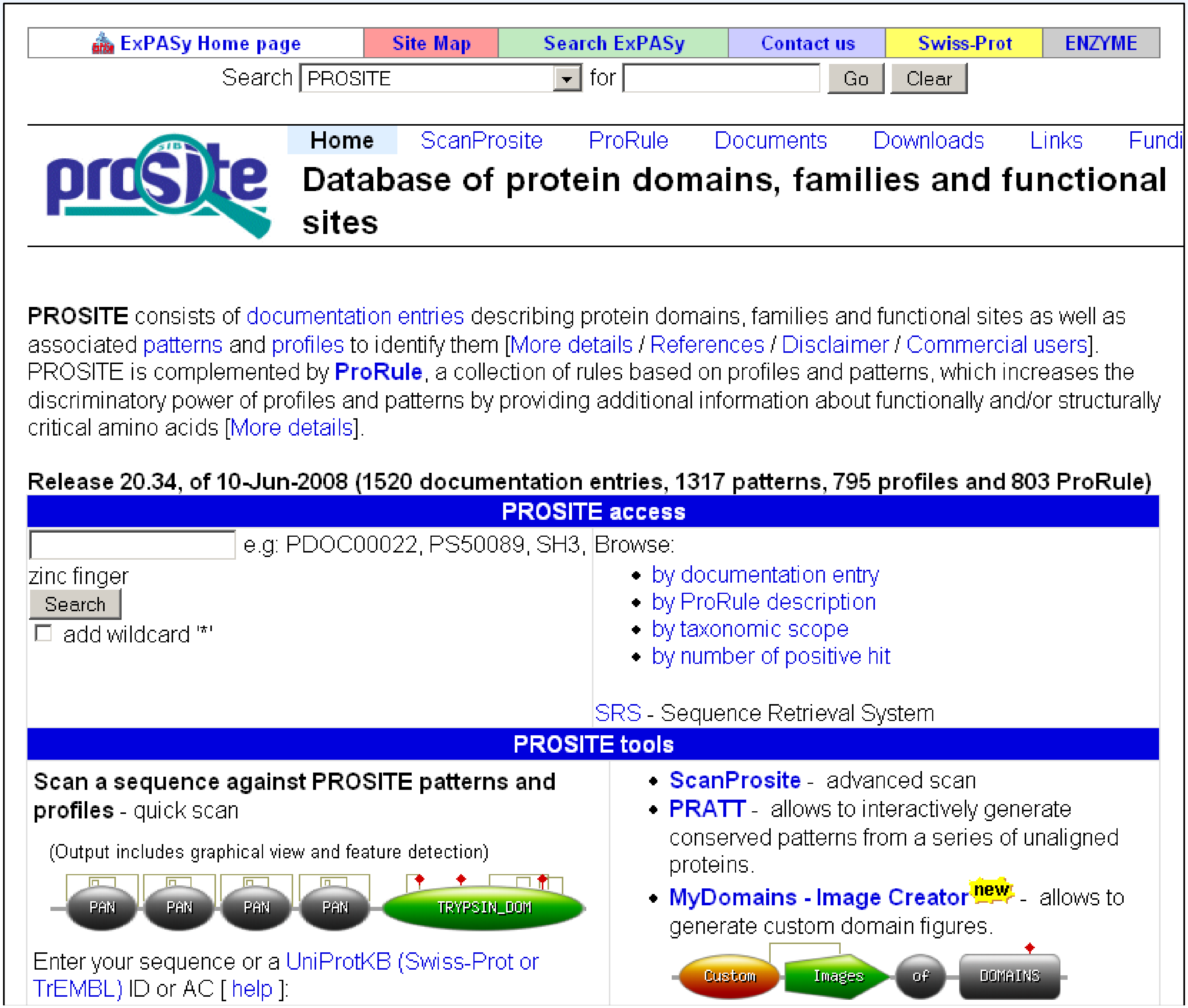
O43325 LYR motif-containing protein 1

O46098 Protein bcn92

P30643 Uncharacterized protein R08D7.4

Prosite

http://ca.expasy.org/prosite/



Database of protein domains, families and functional sites

PROSITE consists of documentation entries describing protein domains, families and functional sites as well as associated patterns and profiles to identify them [More details / References / Disclaimer / Commercial users]. PROSITE is complemented by **ProRule**, a collection of rules based on profiles and patterns, which increases the discriminatory power of profiles and patterns by providing additional information about functionally and/or structurally critical amino acids [More details].

Release 20.34, of 10-Jun-2008 (1520 documentation entries, 1317 patterns, 795 profiles and 803 ProRule)

PROSITE access

e.g: PDOC00022, PS50089, SH3, zinc finger

Search

□ add wildcard '*'

Browse:

- by documentation entry
- by ProRule description
- by taxonomic scope
- by number of positive hit

SRS - Sequence Retrieval System

PROSITE tools

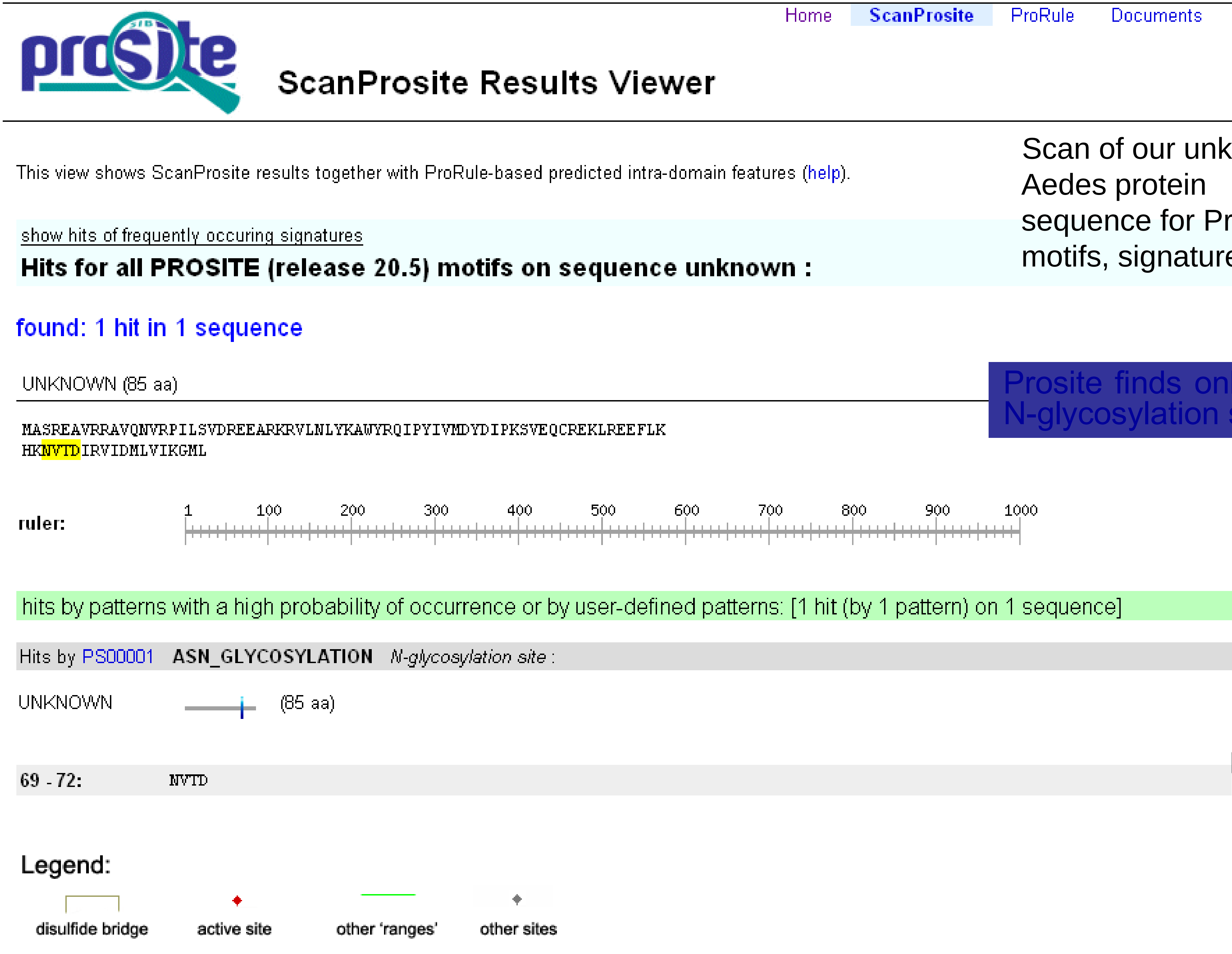
Scan a sequence against PROSITE patterns and profiles - quick scan

(Output includes graphical view and feature detection)

Enter your sequence or a UniProtKB (Swiss-Prot or TrEMBL) ID or AC [help]:

- **ScanProsite** - advanced scan
- **PRATT** - allows to interactively generate conserved patterns from a series of unaligned proteins.
- **MyDomains - Image Creator** new - allows to generate custom domain figures.

Prosite hit for unknown protein



Home | ScanProsite | ProRule | Documents | Downloads | Links

ScanProsite Results Viewer

This view shows ScanProsite results together with ProRule-based predicted intra-domain features (help).

show hits of frequently occuring signatures

Hits for all PROSITE (release 20.5) motifs on sequence unknown :

found: 1 hit in 1 sequence

UNKNOWN (85 aa)

```
MASREAVRRAVQNVRPILSVDREEARKRVLNLYKAWYRQIPYIVMDYDIPKSVEQCREKLREEFLK
HKNVTDIRVIDMLVIKGML
```

ruler: 1 100 200 300 400 500 600 700 800 900 1000

hits by patterns with a high probability of occurrence or by user-defined patterns: [1 hit (by 1 pattern) on 1 sequence]

Hits by PS00001 **ASN_GLYCOSYLATION** *N-glycosylation site* :

UNKNOWN (85 aa)

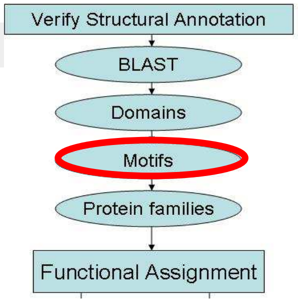
69 - 72: NVTD

Legend:

disulfide bridge | active site | other 'ranges' | other sites

Scan of our unknown Aedes protein sequence for Prosite motifs, signatures

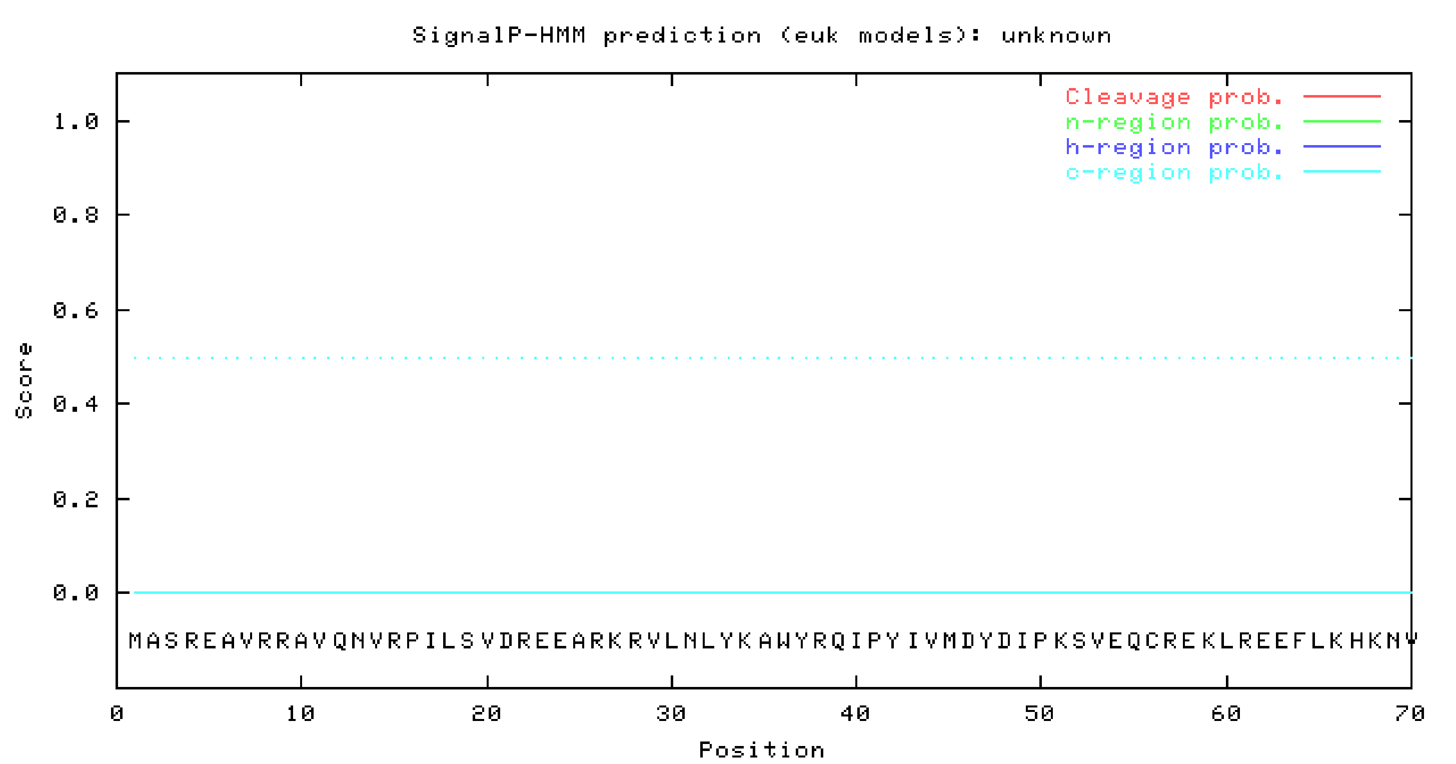
Prosite finds only an N-glycosylation site.



SignalP results

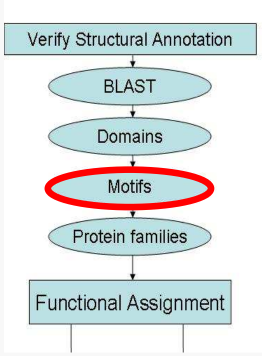
There is no signal sequence in our unknown *Aedes aegypti* protein sequence.

SignalP-HMM result:

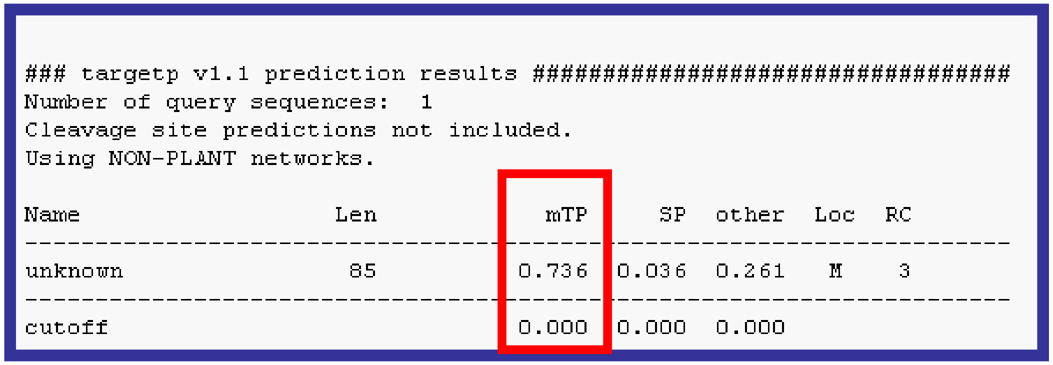


data

```
>unknown
Prediction: Non-secretory protein
Signal peptide probability: 0.000
Signal anchor probability: 0.000
Max cleavage site probability: 0.000 between pos. 19 and 20
```

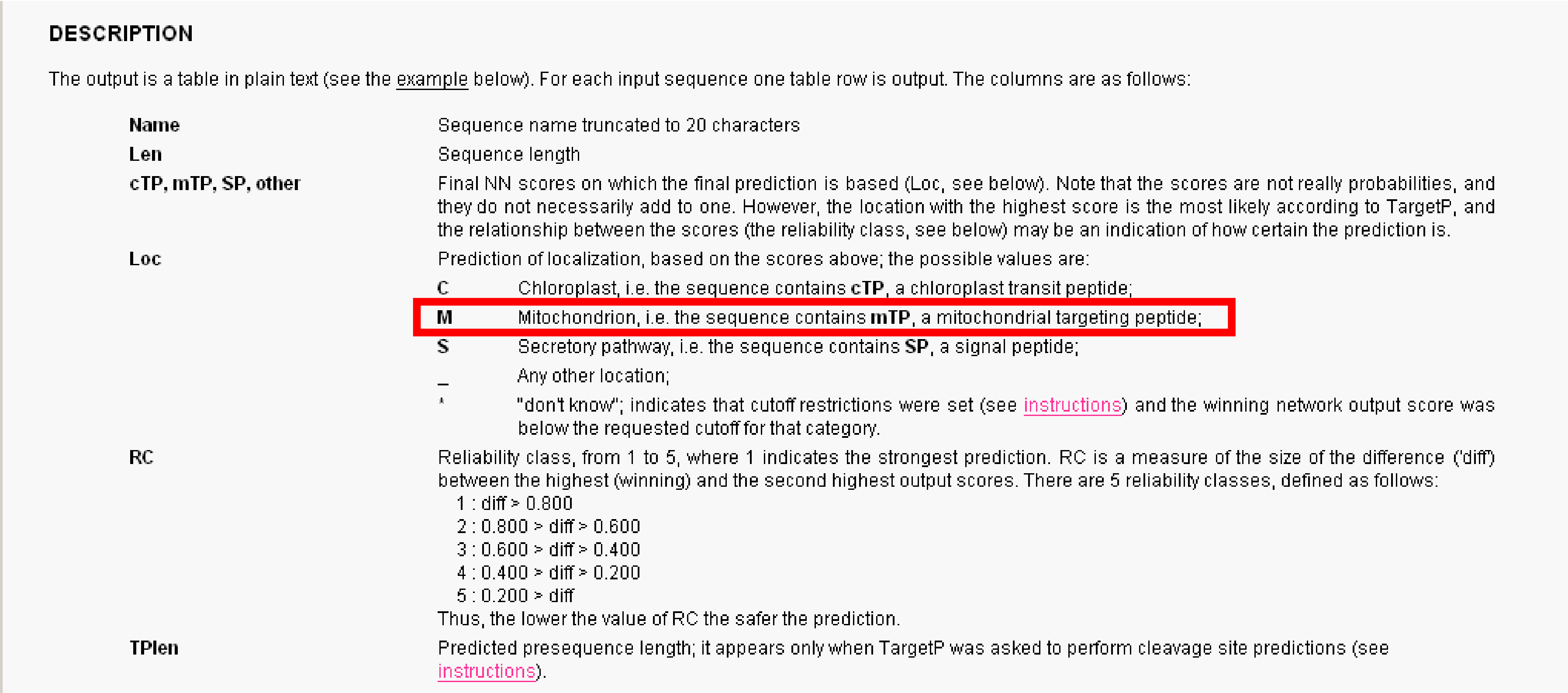


TargetP results



```
### targetp v1.1 prediction results ##################################
Number of query sequences:  1
Cleavage site predictions not included.
Using NON-PLANT networks.

Name                  Len            mTP     SP  other  Loc  RC
----------------------------------------------------------------------
unknown                85          0.736  0.036  0.261   M     3
----------------------------------------------------------------------
cutoff                             0.000  0.000  0.000
```

There is a high probability that our unknown *Aedes aegypti* sequence is targeted to the mitochondrion.



DESCRIPTION

The output is a table in plain text (see the example below). For each input sequence one table row is output. The columns are as follows:

Name Sequence name truncated to 20 characters

Len Sequence length

cTP, mTP, SP, other Final NN scores on which the final prediction is based (Loc, see below). Note that the scores are not really probabilities, and they do not necessarily add to one. However, the location with the highest score is the most likely according to TargetP, and the relationship between the scores (the reliability class, see below) may be an indication of how certain the prediction is.

Loc Prediction of localization, based on the scores above; the possible values are:

C Chloroplast, i.e. the sequence contains **cTP**, a chloroplast transit peptide;

M Mitochondrion, i.e. the sequence contains **mTP**, a mitochondrial targeting peptide;

S Secretory pathway, i.e. the sequence contains **SP**, a signal peptide;

_ Any other location;

* "don't know"; indicates that cutoff restrictions were set (see instructions) and the winning network output score was below the requested cutoff for that category.

RC Reliability class, from 1 to 5, where 1 indicates the strongest prediction. RC is a measure of the size of the difference ('diff') between the highest (winning) and the second highest output scores. There are 5 reliability classes, defined as follows:

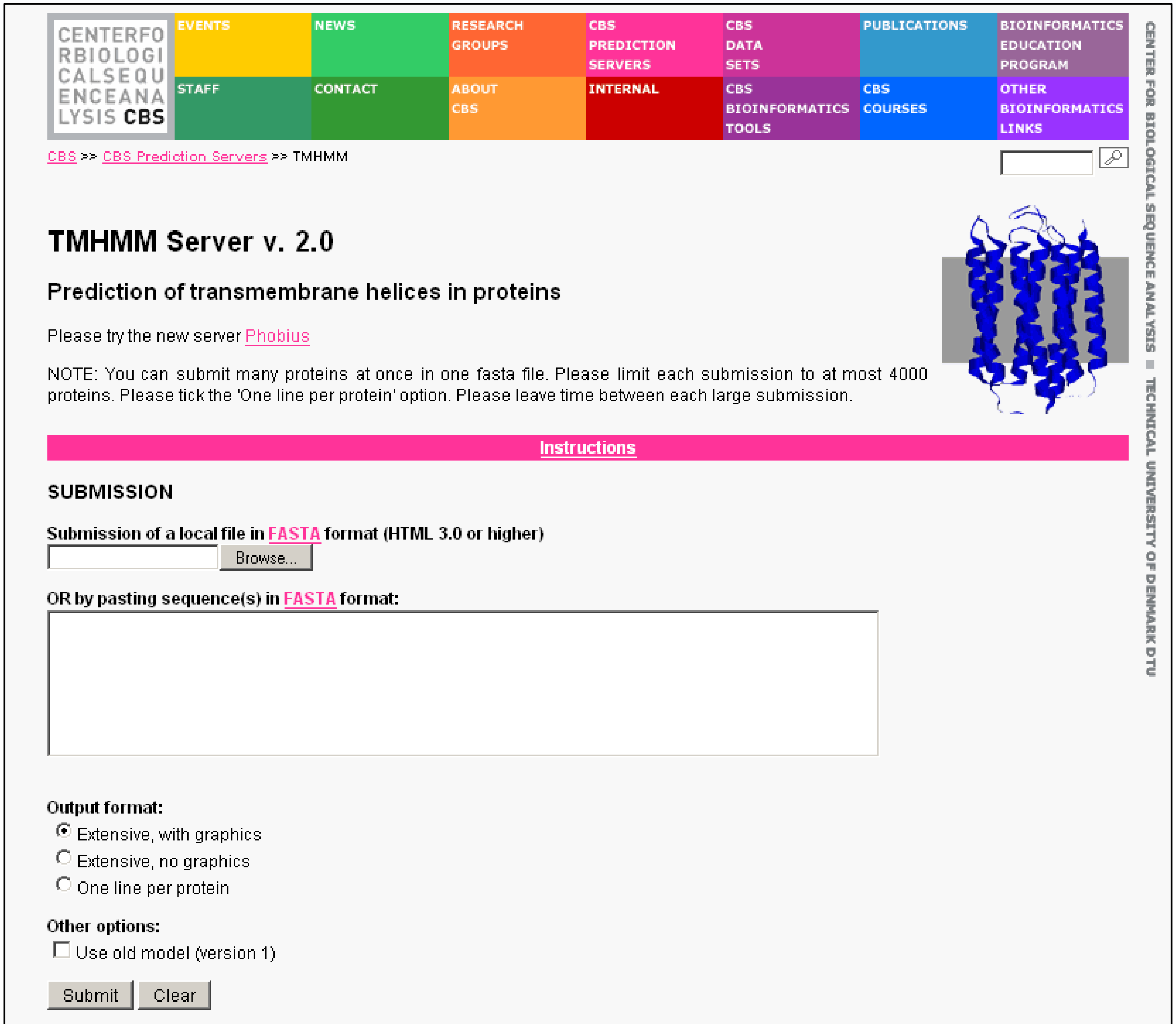
1 : diff > 0.800
2 : 0.800 > diff > 0.600
3 : 0.600 > diff > 0.400
4 : 0.400 > diff > 0.200
5 : 0.200 > diff

Thus, the lower the value of RC the safer the prediction.

TPlen Predicted presequence length; it appears only when TargetP was asked to perform cleavage site predictions (see instructions).

TMHMM

http://www.cbs.dtu.dk/services/TMHMM/


CENTER FOR BIOLOGICAL SEQUENCE ANALYSIS CBS
EVENTS
NEWS
RESEARCH GROUPS
CBS PREDICTION SERVERS
CBS DATA SETS
PUBLICATIONS
BIOINFORMATICS EDUCATION PROGRAM
STAFF
CONTACT
ABOUT CBS
INTERNAL
CBS BIOINFORMATICS TOOLS
CBS COURSES
OTHER BIOINFORMATICS LINKS
CBS >> CBS Prediction Servers >> TMHMM
CENTER FOR BIOLOGICAL SEQUENCE ANALYSIS ■ TECHNICAL UNIVERSITY OF DENMARK DTU

TMHMM Server v. 2.0

Prediction of transmembrane helices in proteins

Please try the new server Phobius

NOTE: You can submit many proteins at once in one fasta file. Please limit each submission to at most 4000 proteins. Please tick the 'One line per protein' option. Please leave time between each large submission.

SUBMISSION

Submission of a local file in FASTA format (HTML 3.0 or higher)

OR by pasting sequence(s) in FASTA format:

Output format:

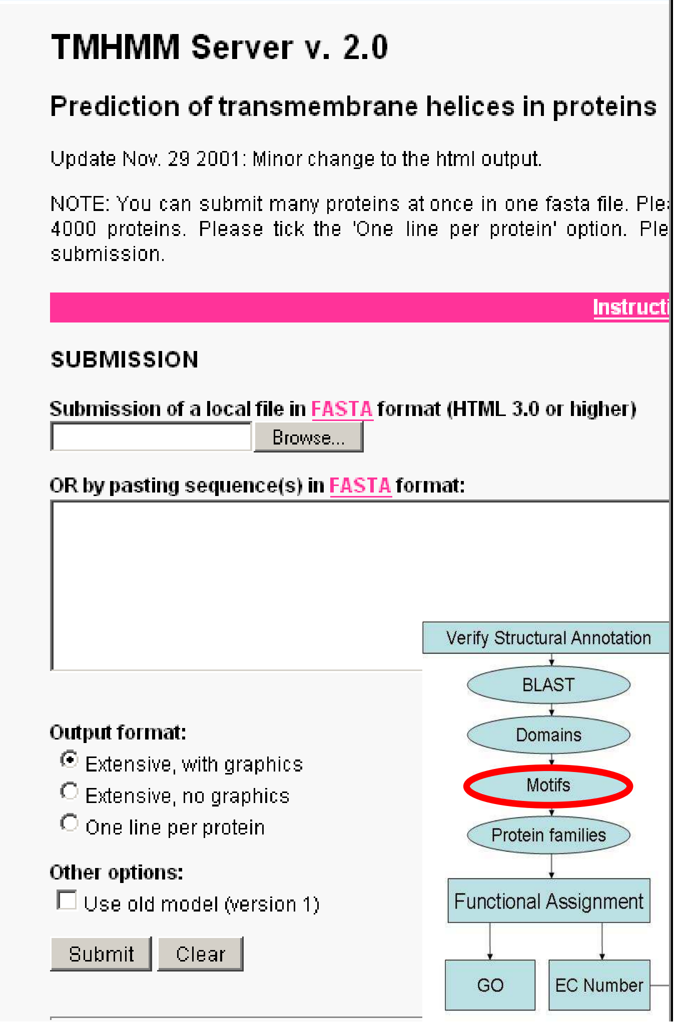
● Extensive, with graphics
○ Extensive, no graphics
○ One line per protein

Other options:

□ Use old model (version 1)

TMHMM – Transmembrane Domain

Our sequence is predicted to have 2 transmembrane domains.



TMHMM Server v. 2.0

Prediction of transmembrane helices in proteins

Update Nov. 29 2001: Minor change to the html output.

NOTE: You can submit many proteins at once in one fasta file. Ple
4000 proteins. Please tick the 'One line per protein' option. Ple
submission.

SUBMISSION

Submission of a local file in FASTA format (HTML 3.0 or higher)

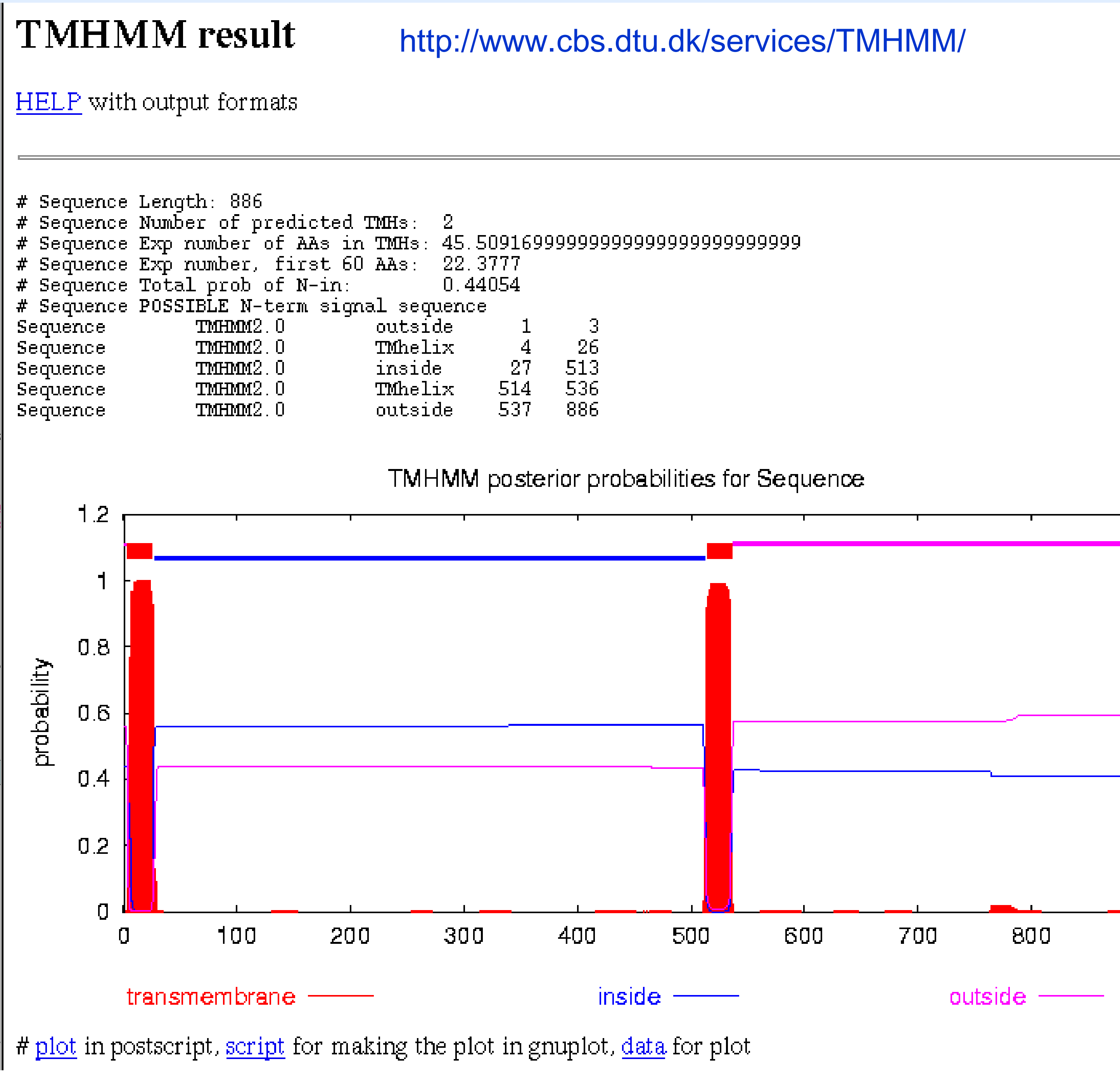
OR by pasting sequence(s) in FASTA format:

Output format:

● Extensive, with graphics
○ Extensive, no graphics
○ One line per protein

Other options:

□ Use old model (version 1)



TMHMM result

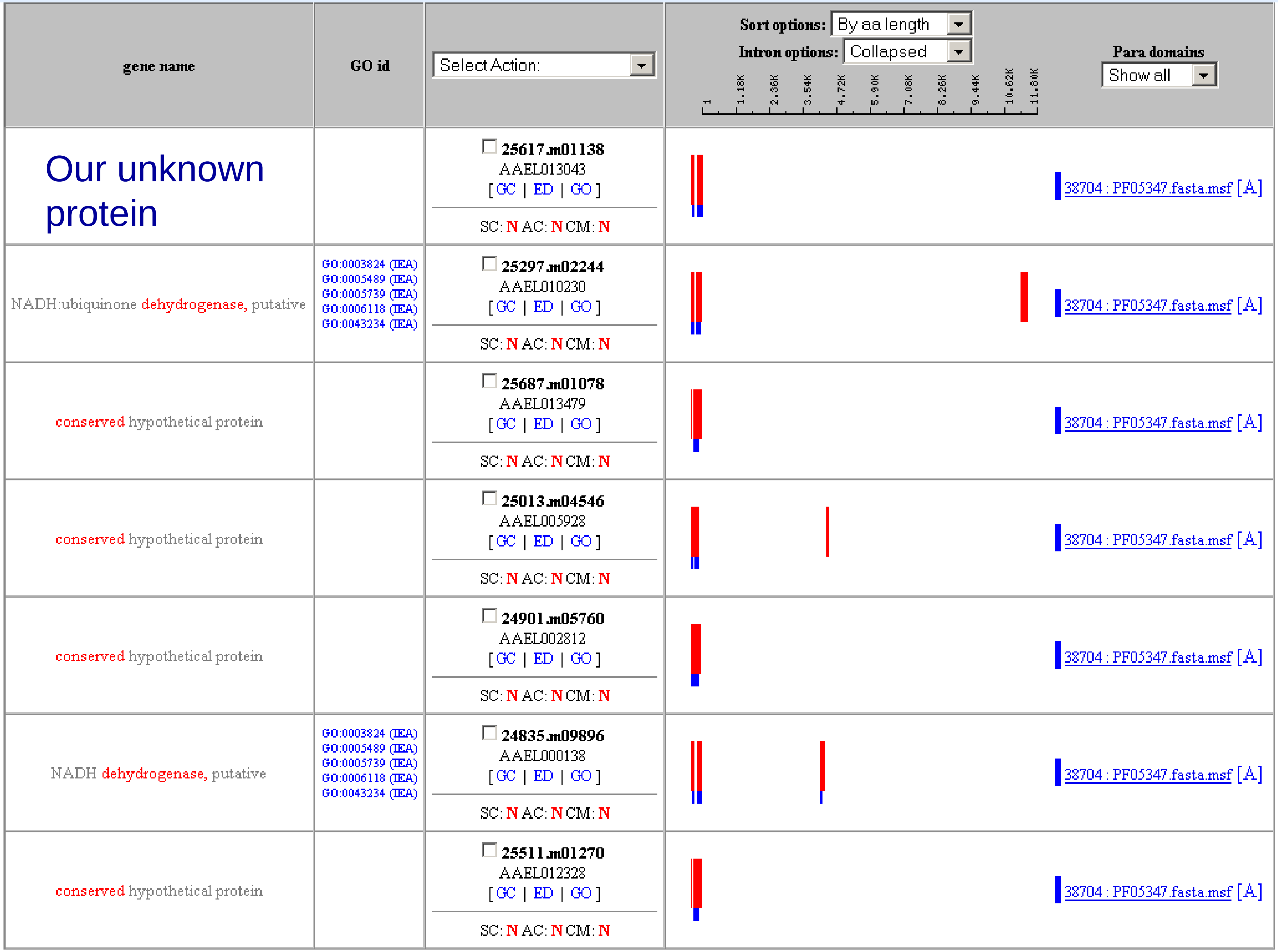
http://www.cbs.dtu.dk/services/TMHMM/

HELP with output formats

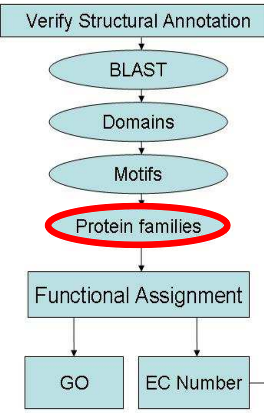
```
# Sequence Length: 886
# Sequence Number of predicted TMHs:  2
# Sequence Exp number of AAs in TMHs: 45.509169999999999999999999999
# Sequence Exp number, first 60 AAs:  22.3777
# Sequence Total prob of N-in:        0.44054
# Sequence POSSIBLE N-term signal sequence
Sequence        TMHMM2.0        outside      1     3
Sequence        TMHMM2.0        TMhelix      4    26
Sequence        TMHMM2.0        inside      27   513
Sequence        TMHMM2.0        TMhelix    514   536
Sequence        TMHMM2.0        outside    537   886
```

plot in postscript, script for making the plot in gnuplot, data for plot

JCVI Paralogous Families

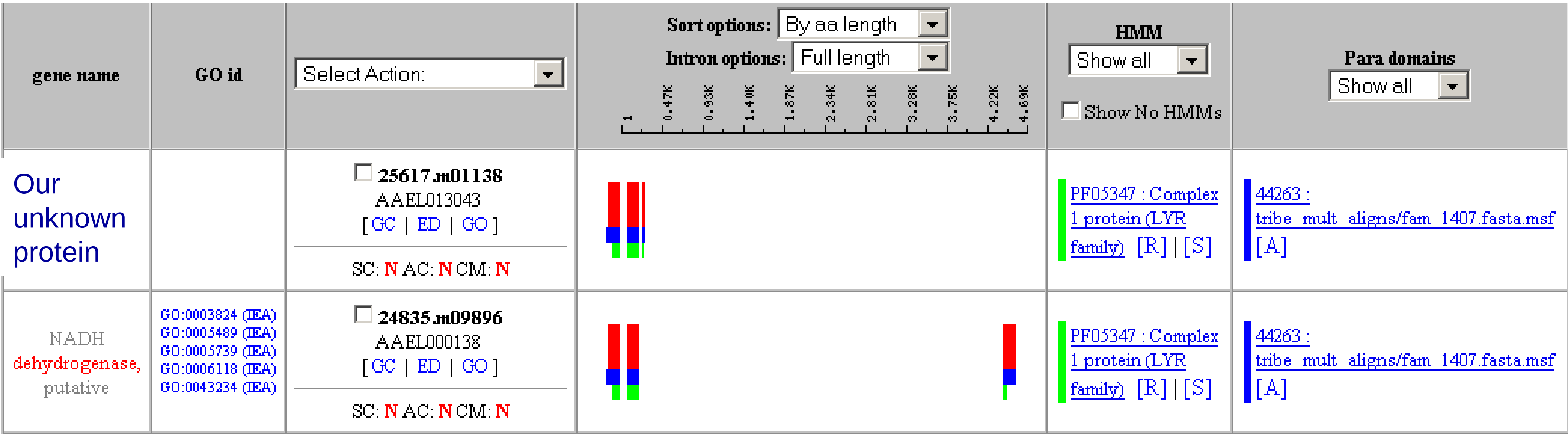


gene name	GO id	Select Action: ▾	Sort options: By aa length ▾ Intron options: Collapsed ▾ 1 1.18K 2.36K 3.54K 4.72K 5.90K 7.08K 8.26K 9.44K 10.62K 11.80K	Para domains Show all ▾
Our unknown protein		□ 25617.m01138 AAEL013043 [GC | ED | GO] SC: N AC: N CM: N		38704 : PF05347.fasta.msf [A]
NADH:ubiquinone dehydrogenase, putative	GO:0003824 (IEA) GO:0005489 (IEA) GO:0005739 (IEA) GO:0006118 (IEA) GO:0043234 (IEA)	□ 25297.m02244 AAEL010230 [GC | ED | GO] SC: N AC: N CM: N		38704 : PF05347.fasta.msf [A]
conserved hypothetical protein		□ 25687.m01078 AAEL013479 [GC | ED | GO] SC: N AC: N CM: N		38704 : PF05347.fasta.msf [A]
conserved hypothetical protein		□ 25013.m04546 AAEL005928 [GC | ED | GO] SC: N AC: N CM: N		38704 : PF05347.fasta.msf [A]
conserved hypothetical protein		□ 24901.m05760 AAEL002812 [GC | ED | GO] SC: N AC: N CM: N		38704 : PF05347.fasta.msf [A]
NADH dehydrogenase, putative	GO:0003824 (IEA) GO:0005489 (IEA) GO:0005739 (IEA) GO:0006118 (IEA) GO:0043234 (IEA)	□ 24835.m09896 AAEL000138 [GC | ED | GO] SC: N AC: N CM: N		38704 : PF05347.fasta.msf [A]
conserved hypothetical protein		□ 25511.m01270 AAEL012328 [GC | ED | GO] SC: N AC: N CM: N		38704 : PF05347.fasta.msf [A]

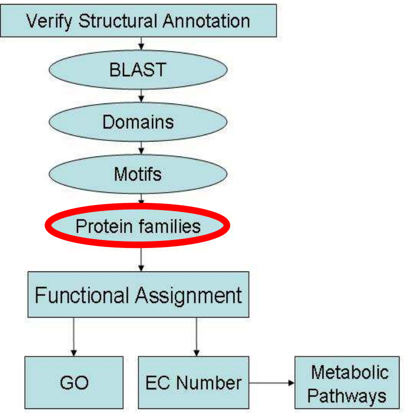


TribeMCL

http://www.ebi.ac.uk/research/cgg/tribe/



gene name	GO id	Select Action:	Sort options: By aa length Intron options: Full length 1 0.47K 0.93K 1.40K 1.87K 2.34K 2.81K 3.28K 3.75K 4.22K 4.69K	HMM Show all ☐ Show No HMMs	Para domains Show all
Our unknown protein		☐ 25617.m01138 AAEL013043 [GC | ED | GO] SC: N AC: N CM: N		PF05347 : Complex 1 protein (LYR family) [R] | [S]	44263 : tribe_mult_aligns/fam_1407.fasta.msf [A]
NADH dehydrogenase, putative	GO:0003824 (IEA) GO:0005489 (IEA) GO:0005739 (IEA) GO:0006118 (IEA) GO:0043234 (IEA)	☐ 24835.m09896 AAEL000138 [GC | ED | GO] SC: N AC: N CM: N		PF05347 : Complex 1 protein (LYR family) [R] | [S]	44263 : tribe_mult_aligns/fam_1407.fasta.msf [A]



Verify Structural Annotation

BLAST

Domains

Motifs

Protein families

Functional Assignment

GO

EC Number

Metabolic Pathways

An Overview of Similarity Search Results #1

BLAST: similarity to many NADH-ubiquinone oxidoreductases, and one significant hit to an experimentally characterized protein: NADH dehydrogenase (ubiquinone) 1 alpha subcomplex, 6, 14kDa [Homo sapiens] .

Domain: PF05347: **Complex 1 protein (LYR family)** Good alignment to seed.

Total score: **72.5** Trusted cutoff: **25.00** Noise cutoff: **24.40** Total expect: **1.5e-18**

Proteins in this family have been identified as a component of the higher eukaryotic NADH complex and may play a role in Fe/S cluster biogenesis in mitochondria.. In Saccharomyces cerevisiae, the Isd11 protein (Q6Q560_YEAST) has been shown to play a role in Fe/S cluster biogenesis in mitochondria. *The family includes proteins from the NADH-ubiquinone oxidoreductase complex I.*

Interpro: Complex 1 LYR protein family

This family of short proteins includes proteins from the NADH-ubiquinone oxidoreductase complex I.

An Overview of Similarity Search Results #2

- **Prosite** scan found one N-glycosylation site.
- **SignalP:** no signal sequence found.
- **TargetP**: There is a high probability that our unknown *Aedes aegypti* sequence is targeted to the mitochondrion.
- **TmHMM:** The sequence contains 2 probable transmembrane domains.
- **Protein Families:** Inconclusive, but not inconsistent. TIGR Paralogous families has sequence as a member of a family containing two “putative” NADH dehydrogenases and four “conserved hypothetical” proteins. None of the family members are characterized. It is a member of a TribeMCL cluster with one “putative” NADH dehydrogenase, which is not characterized.

High Confidence Naming

To have high-confidence in precise function,

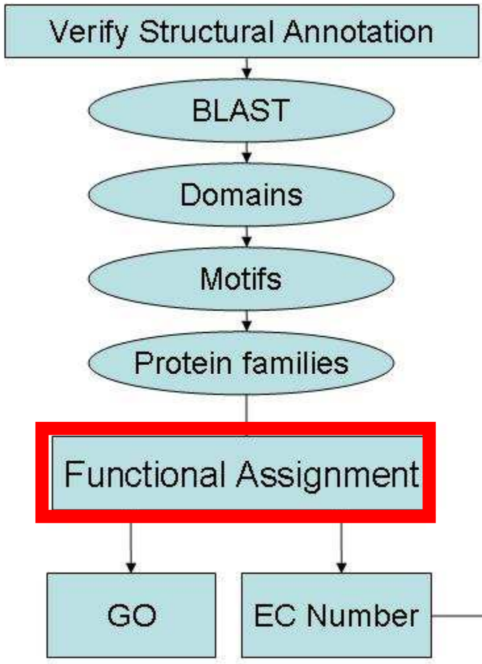
you must have:

- At least one good alignment to an **experimentally characterized** protein
- Hits to HMM Above the Trusted Cutoff
- Conserved active sites, binding sites, appropriate number of membrane spans, etc.
- If no evidence, name it “hypothetical protein”

Example 2: Functional assignment

We have a choice of naming this protein after the domain, “LYR motif family protein” or “LYR motif-containing protein, or we could name it after the human NADH dehydrogenase (ubiquinone) 1 alpha subcomplex, 6 protein. However, to have confidence that our protein MIGHT have the same function, we would need better than a 41% match. One option would be to call it “NADH dehydrogenase (ubiquinone) subunit, putative.”

Our curator might call it “LYR motif family protein” – or “hypothetical protein.”



Curation Input via Manatee

Gene name

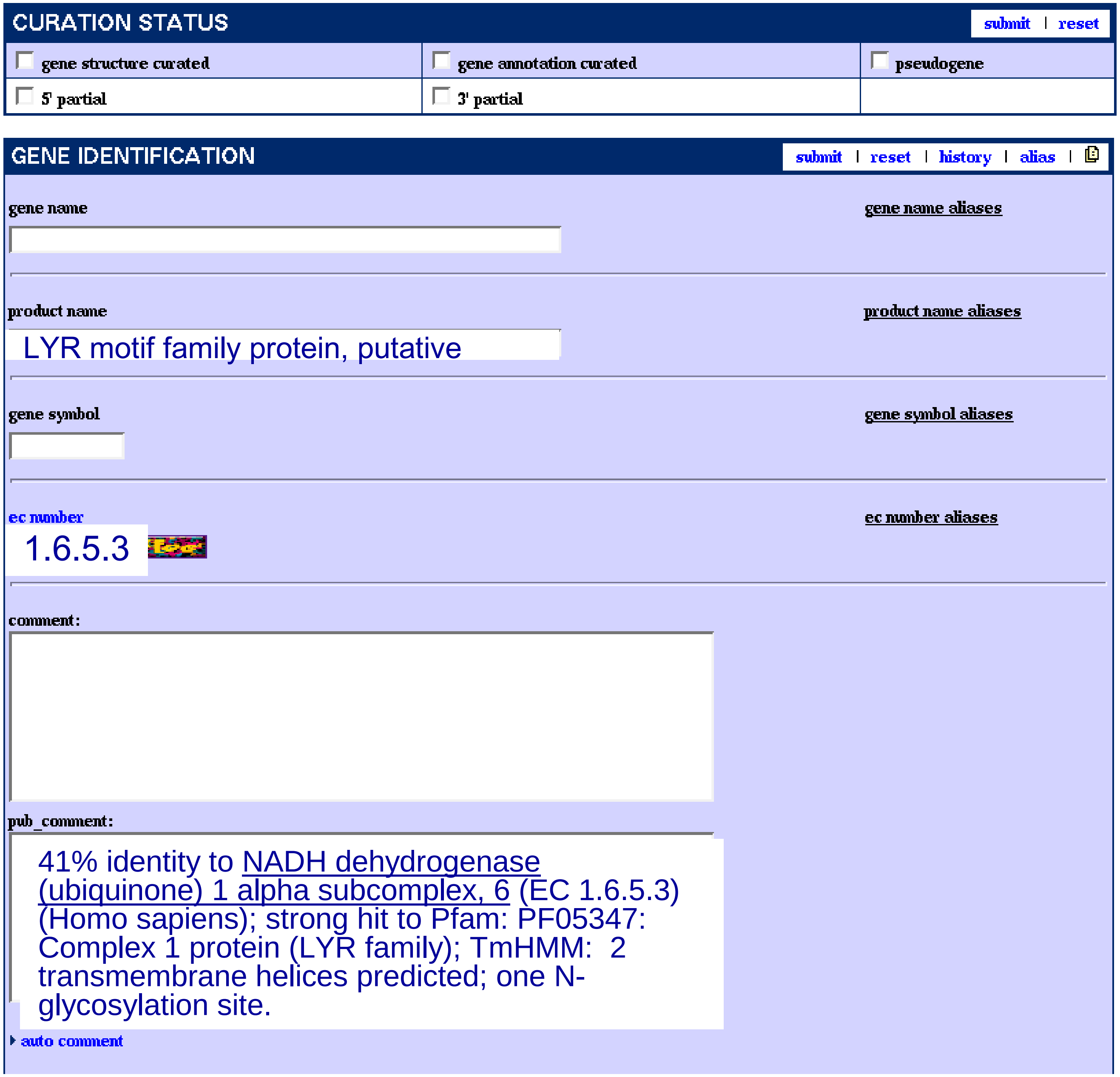
Gene product name

Gene symbol

EC number

Internal coments

Public comments



CURATION STATUS

□ gene structure curated	□ gene annotation curated	□ pseudogene
□ 5' partial	□ 3' partial	

GENE IDENTIFICATION

gene name

gene name aliases

product name: LYR motif family protein, putative

product name aliases

gene symbol

gene symbol aliases

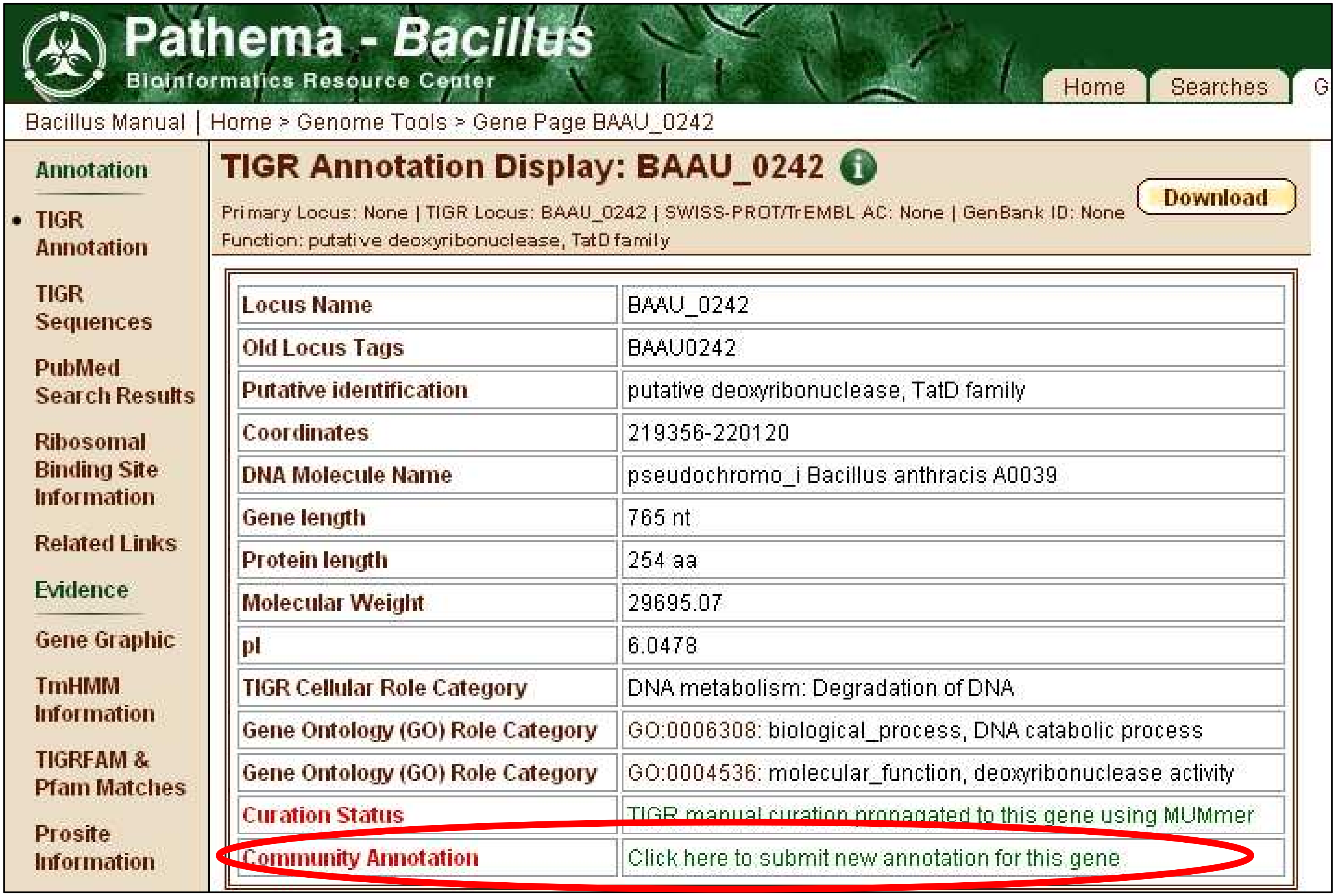
ec number: 1.6.5.3

ec number aliases

comment:

pub_comment: 41% identity to NADH dehydrogenase (ubiquinone) 1 alpha subcomplex, 6 (EC 1.6.5.3) (Homo sapiens); strong hit to Pfam: PF05347: Complex 1 protein (LYR family); TmHMM: 2 transmembrane helices predicted; one N-glycosylation site.

Community Annotation



Bacillus Manual | Home > Genome Tools > Gene Page BAAU_0242

Annotation

- TIGR Annotation

TIGR Sequences

PubMed Search Results

Ribosomal Binding Site Information

Related Links

Evidence

Gene Graphic

TmHMM Information

TIGRFAM & Pfam Matches

Prosite Information

TIGR Annotation Display: BAAU_0242

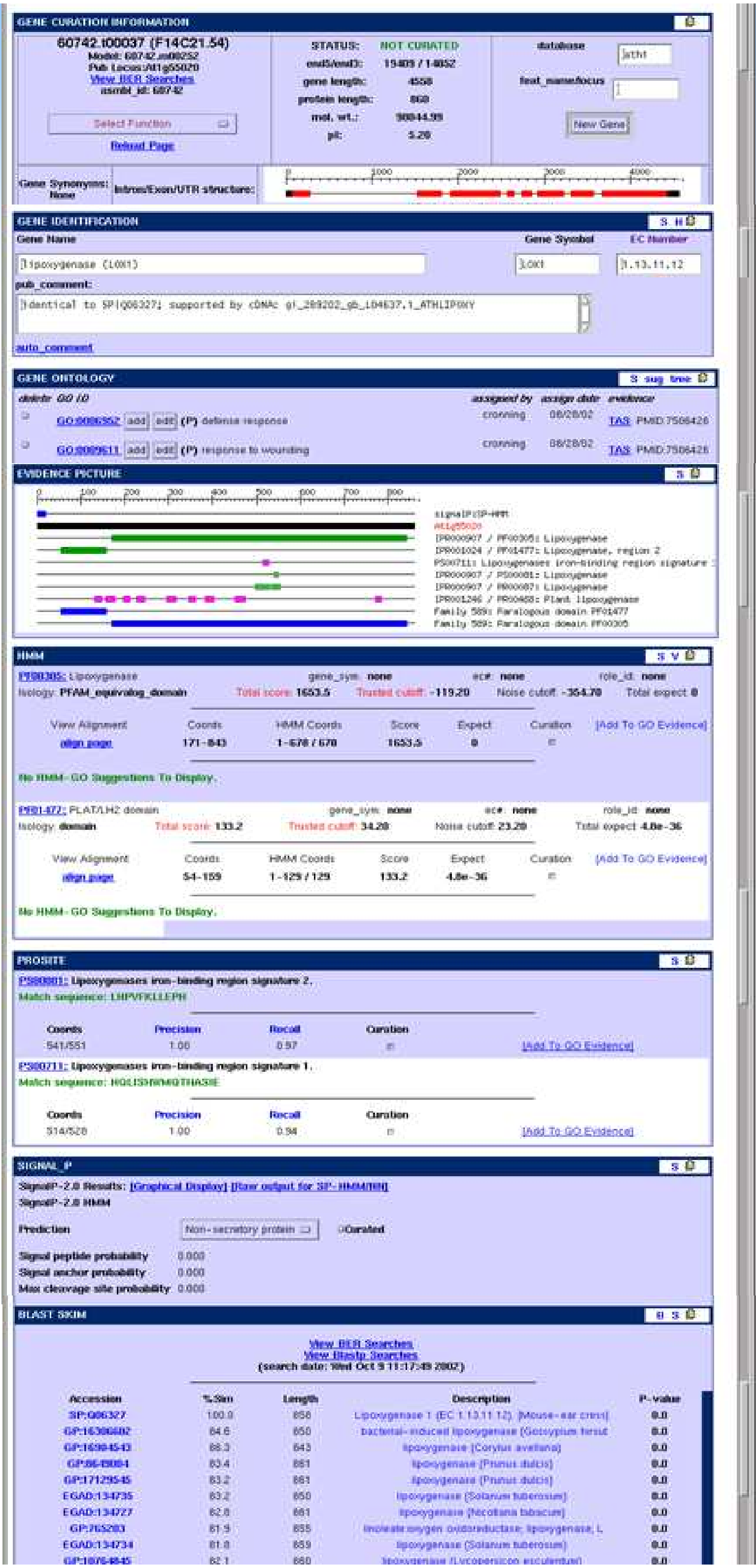
Primary Locus: None | TIGR Locus: BAAU_0242 | SWISS-PROT/TrEMBL AC: None | GenBank ID: None

Function: putative deoxyribonuclease, TatD family

Locus Name	BAAU_0242
Old Locus Tags	BAAU0242
Putative identification	putative deoxyribonuclease, TatD family
Coordinates	219356-220120
DNA Molecule Name	pseudochromo_i Bacillus anthracis A0039
Gene length	765 nt
Protein length	254 aa
Molecular Weight	29695.07
pI	6.0478
TIGR Cellular Role Category	DNA metabolism: Degradation of DNA
Gene Ontology (GO) Role Category	GO:0006308: biological_process, DNA catabolic process
Gene Ontology (GO) Role Category	GO:0004536: molecular_function, deoxyribonuclease activity
Curation Status	TIGR manual curation propagated to this gene using MUMmer
Community Annotation	Click here to submit new annotation for this gene

MANATEE

- All of the searches shown are available in a Manatee installation, with a database and computational pipeline.
- Navigation, inspection & <u>curation</u> of gene products
  - Gene/Gene products
  - GO Assignments
- Available at:
  - http://manatee.sourceforge.net



GENE CURATION INFORMATION

60742.t00037 (F14C21.54)
Model: 60742.m00252
Pub Locus:At1g55020
View BER Searches
asmbl_id: 60742

Select Function
Reload Page

STATUS:	NOT CURATED
end5/end3:	15409 / 14852
gene length:	4558
protein length:	860
mol. wt.:	98044.99
pI:	5.29

database: atht
feat_name/locus:
New Gene

Gene Synonyms: None
Intron/Exon/UTR structure:

GENE IDENTIFICATION

Gene Name	Gene Symbol	EC Number
lipoxygenase (LOX1)	LOX1	1.13.11.12

pub_comment:
Identical to SP|Q06327; supported by cDNA: gi_289202_gb_L04637.1_ATHLIPOXY

auto_comment

GENE ONTOLOGY

delete	GO ID				assigned by	assign date	evidence
□	GO:0006952	add	edit	(P) defense response	cronning	08/28/02	TAS PMID:7506426
□	GO:0009611	add	edit	(P) response to wounding	cronning	08/28/02	TAS PMID:7506426

EVIDENCE PICTURE

signalP:SP-HMM
At1g55020
IPR000907 / PF00305: Lipoxygenase
IPR001024 / PF01477: Lipoxygenase, region 2
PS00711: Lipoxygenases iron-binding region signature 1
IPR000907 / PS00081: Lipoxygenase
IPR000907 / PR00087: Lipoxygenase
IPR001246 / PR00468: Plant lipoxygenase
Family 589: Paralogous domain PF01477
Family 589: Paralogous domain PF00305

HMM

PF00305: Lipoxygenase gene_sym: none ec#: none role_id: none
Isology: PFAM_equivalog_domain Total score: 1653.5 Trusted cutoff: -119.20 Noise cutoff: -354.70 Total expect: 0

View Alignment	Coords	HMM Coords	Score	Expect	Curation	[Add To GO Evidence]
align page	171-843	1-678 / 678	1653.5	0	□	

No HMM-GO Suggestions To Display.

PF01477: PLAT/LH2 domain gene_sym: none ec#: none role_id: none
Isology: domain Total score: 133.2 Trusted cutoff: 34.20 Noise cutoff: 23.20 Total expect: 4.8e-36

View Alignment	Coords	HMM Coords	Score	Expect	Curation	[Add To GO Evidence]
align page	54-159	1-129 / 129	133.2	4.8e-36	□	

No HMM-GO Suggestions To Display.

PROSITE

PS00081: Lipoxygenases iron-binding region signature 2.
Match sequence: LHPVFKLLEPH

Coords	Precision	Recall	Curation	
541/551	1.00	0.97	□	[Add To GO Evidence]

PS00711: Lipoxygenases iron-binding region signature 1.
Match sequence: HQLISHWMQTHASIE

Coords	Precision	Recall	Curation	
514/528	1.00	0.94	□	[Add To GO Evidence]

SIGNAL_P

SignalP-2.0 Results: [Graphical Display] [Raw output for SP-HMM/NN]
SignalP-2.0 HMM

Prediction Non-secretory protein □Curated

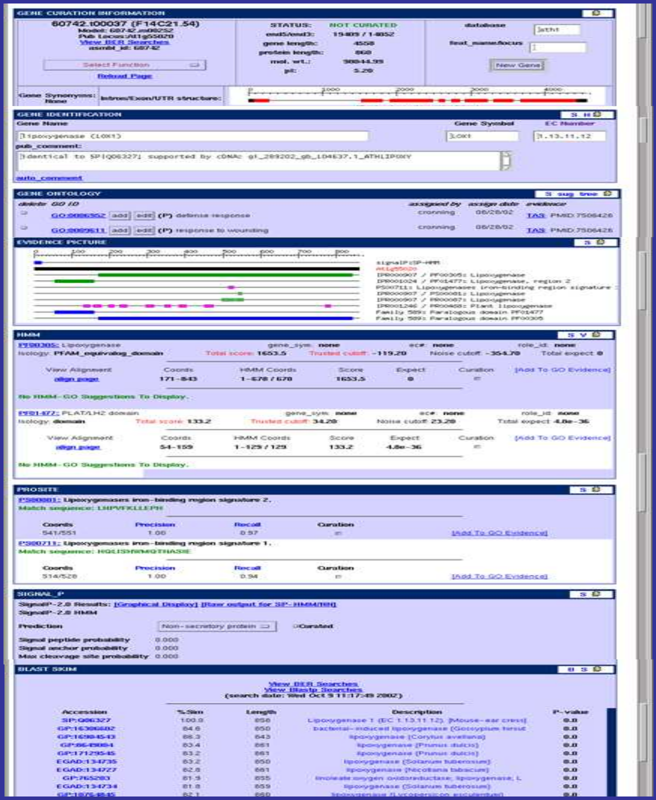
Signal peptide probability	0.000
Signal anchor probability	0.000
Max cleavage site probability	0.000

BLAST SKIM

View BER Searches
View Blastp Searches
(search date: Wed Oct 9 11:17:49 2002)

Accession	%Sim	Length	Description	P-value
SP:Q06327	100.0	856	Lipoxygenase 1 (EC 1.13.11.12) [Mouse-ear cress]	0.0
GP:16306602	84.6	850	bacterial-induced lipoxygenase (Gossypium hirsut	0.0
GP:16904543	88.3	843	lipoxygenase (Corylus avellana)	0.0
GP:8649004	83.4	861	lipoxygenase (Prunus dulcis)	0.0
GP:17129545	83.2	861	lipoxygenase (Prunus dulcis)	0.0
EGAD:134735	83.2	850	lipoxygenase (Solanum tuberosum)	0.0
EGAD:134727	82.8	861	lipoxygenase (Nicotiana tabacum)	0.0
GP:765203	81.9	855	linoleate:oxygen oxidoreductase; lipoxygenase; L	0.0
EGAD:134734	81.8	859	lipoxygenase (Solanum tuberosum)	0.0
GP:1076845	82.1	860	lipoxygenase (Lycopersicon esculentum)	0.0

Questions?



GENE CURATION INFORMATION

60742.t00037 (F14C21.54)
Model: 68742.m08252
Pub Locus:At1g55020
View BER Searches
asmbl_id: 68742

Select Function
Reload Page

STATUS:	NOT CURATED	
end5/end3:	19409 / 14852	
gene length:	4558	
protein length:	860	
mol. wt.:	98844.99	
pI:	5.29	

database: atht
feat_name/locus:
New Gene

Gene Synonyms: None
Intron/Exon/UTR structure:

GENE IDENTIFICATION

Gene Name	Gene Symbol	EC Number
lipoxygenase (LOX1)	LOX1	1.13.11.12

pub_comment:
identical to SP\|Q06327; supported by cDNA: gi_289202_gb_L04637.1_ATHLIPOXY

auto_comment

GENE ONTOLOGY

delete	GO ID				assigned by	assign date	evidence
○	GO:0006952	add	edit	(P) defense response	cronning	08/28/02	TAS PMID:7506426
○	GO:0009611	add	edit	(P) response to wounding	cronning	08/28/02	TAS PMID:7506426

EVIDENCE PICTURE

signalP:SP-HMM
At1g55020
IPR000907 / PF00305: Lipoxygenase
IPR001024 / PF01477: Lipoxygenase, region 2
PS00711: Lipoxygenases iron-binding region signature
IPR000907 / PS00081: Lipoxygenase
IPR000907 / PR00087: Lipoxygenase
IPR001246 / PR00468: Plant lipoxygenase
Family 589: Paralogous domain PF01477
Family 589: Paralogous domain PF00305

HMM

PF00305: Lipoxygenase gene_sym: none ec#: none role_id: none
Isology: PFAM_equivalog_domain Total score: 1653.5 Trusted cutoff: -119.20 Noise cutoff: -354.70 Total expect: 0

View Alignment	Coords	HMM Coords	Score	Expect	Curation	
align page	171-843	1-678 / 678	1653.5	0	□	[Add To GO Evidence]

No HMM-GO Suggestions To Display.

PF01477: PLAT/LH2 domain gene_sym: none ec#: none role_id: none
Isology: domain Total score: 133.2 Trusted cutoff: 34.20 Noise cutoff: 23.20 Total expect: 4.8e-36

View Alignment	Coords	HMM Coords	Score	Expect	Curation	
align page	54-159	1-129 / 129	133.2	4.8e-36	□	[Add To GO Evidence]

No HMM-GO Suggestions To Display.

PROSITE

PS00081: Lipoxygenases iron-binding region signature 2.
Match sequence: LHPVFKLLEPH

Coords	Precision	Recall	Curation	
541/551	1.00	0.97	□	[Add To GO Evidence]

PS00711: Lipoxygenases iron-binding region signature 1.
Match sequence: HQLISHWMQTHASIE

Coords	Precision	Recall	Curation	
514/528	1.00	0.94	□	[Add To GO Evidence]

SIGNAL_P

SignalP-2.0 Results: [Graphical Display] [Raw output for SP-HMM/NN]
SignalP-2.0 HMM

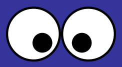
Prediction: Non-secretory protein □Curated

Signal peptide probability	0.000
Signal anchor probability	0.000
Max cleavage site probability	0.000

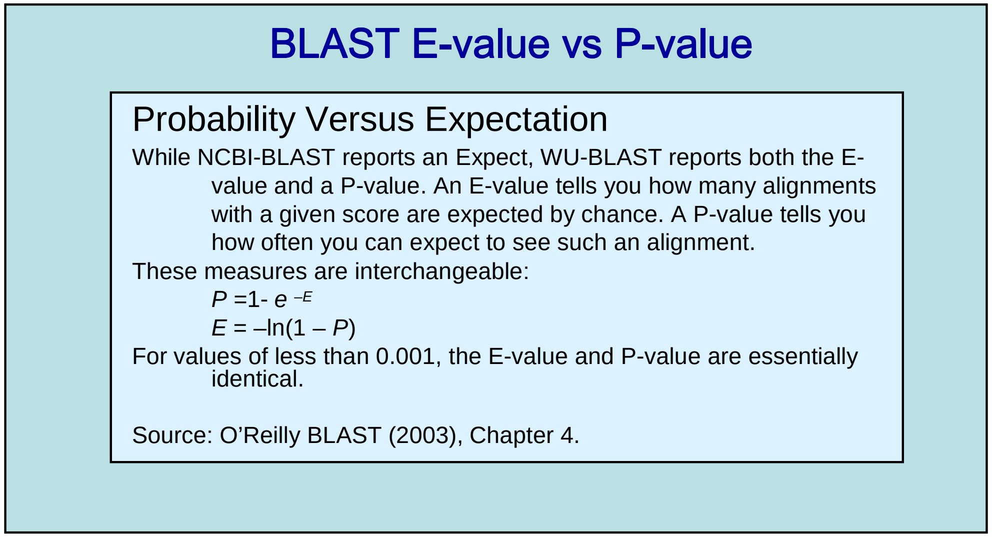
BLAST SKIM

View BER Searches
View Blastp Searches
(search date: Wed Oct 9 11:17:49 2002)

Accession	%Sim	Length	Description	P-value
SP:Q06327	100.0	858	Lipoxygenase 1 (EC 1.13.11.12). [Mouse-ear cress]	0.0
GP:16306602	64.6	850	bacterial-induced lipoxygenase [Gossypium hirsut	0.0
GP:16904543	86.3	843	lipoxygenase [Corylus avellana]	0.0
GP:8649084	83.4	861	lipoxygenase [Prunus dulcis]	0.0
GP:17129545	83.2	861	lipoxygenase [Prunus dulcis]	0.0
EGAD:134735	83.2	850	lipoxygenase [Solanum tuberosum]	0.0
EGAD:134727	82.8	861	lipoxygenase [Nicotiana tabacum]	0.0
GP:765203	81.9	855	linoleate:oxygen oxidoreductase; lipoxygenase; L	0.0
EGAD:134734	81.8	859	lipoxygenase [Solanum tuberosum]	0.0
GP:1076845	82.1	860	lipoxygenase [Lycopersicon esculentum]	0.0



BLAST E-value vs P-value



BLAST E-value vs P-value

Probability Versus Expectation

While NCBI-BLAST reports an Expect, WU-BLAST reports both the E-value and a P-value. An E-value tells you how many alignments with a given score are expected by chance. A P-value tells you how often you can expect to see such an alignment.

These measures are interchangeable:

$P = 1 - e^{-E}$

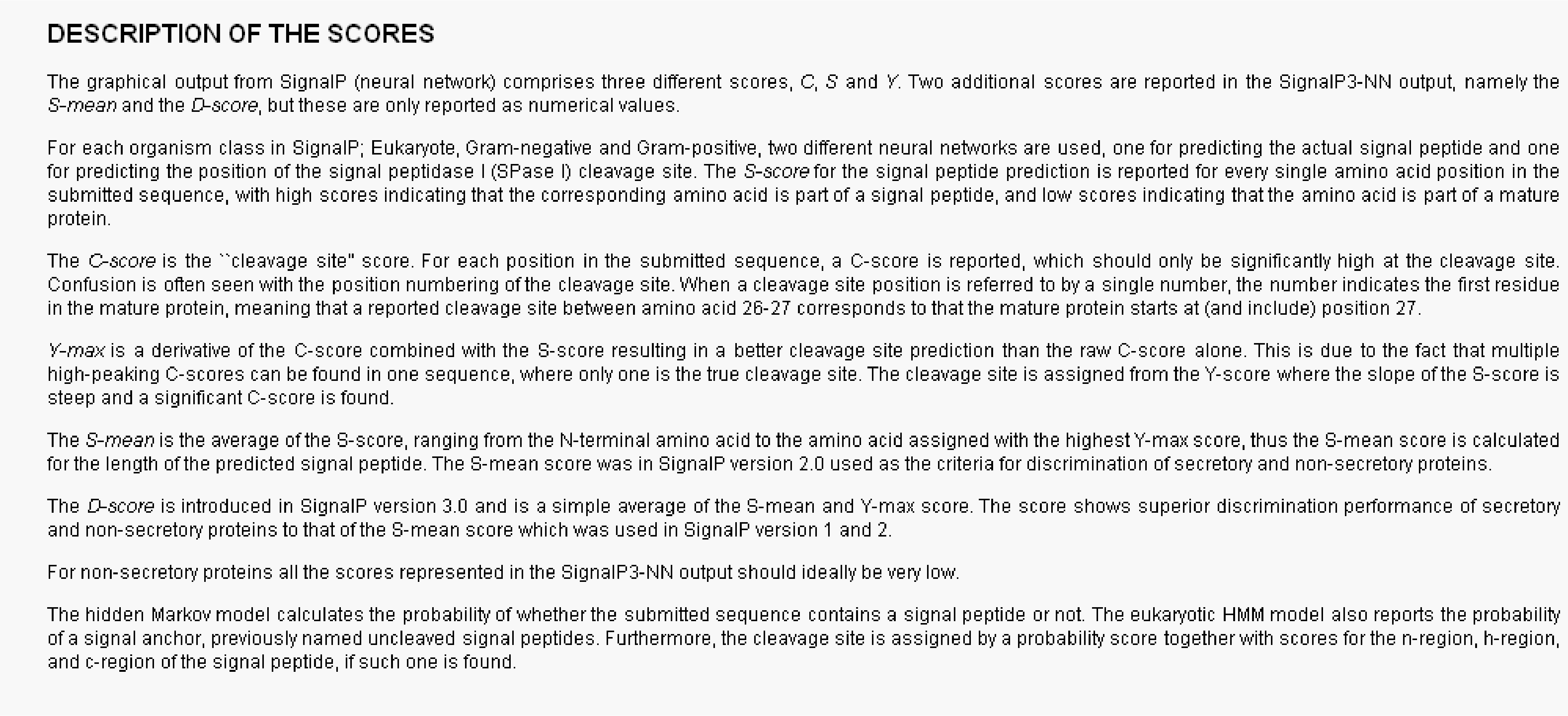
$E = -\ln(1 - P)$

For values of less than 0.001, the E-value and P-value are essentially identical.

Source: O’Reilly BLAST (2003), Chapter 4.

Further Reading:
Ian Korf, Mark Yandell and Joseph Bedell, BLAST, O’Reilly & Associates, Inc., 2003.

SignalP output



DESCRIPTION OF THE SCORES

The graphical output from SignalP (neural network) comprises three different scores, *C*, *S* and *Y*. Two additional scores are reported in the SignalP3-NN output, namely the *S-mean* and the *D-score*, but these are only reported as numerical values.

For each organism class in SignalP; Eukaryote, Gram-negative and Gram-positive, two different neural networks are used, one for predicting the actual signal peptide and one for predicting the position of the signal peptidase I (SPase I) cleavage site. The *S-score* for the signal peptide prediction is reported for every single amino acid position in the submitted sequence, with high scores indicating that the corresponding amino acid is part of a signal peptide, and low scores indicating that the amino acid is part of a mature protein.

The *C-score* is the ``cleavage site" score. For each position in the submitted sequence, a C-score is reported, which should only be significantly high at the cleavage site. Confusion is often seen with the position numbering of the cleavage site. When a cleavage site position is referred to by a single number, the number indicates the first residue in the mature protein, meaning that a reported cleavage site between amino acid 26-27 corresponds to that the mature protein starts at (and include) position 27.

Y-max is a derivative of the C-score combined with the S-score resulting in a better cleavage site prediction than the raw C-score alone. This is due to the fact that multiple high-peaking C-scores can be found in one sequence, where only one is the true cleavage site. The cleavage site is assigned from the Y-score where the slope of the S-score is steep and a significant C-score is found.

The *S-mean* is the average of the S-score, ranging from the N-terminal amino acid to the amino acid assigned with the highest Y-max score, thus the S-mean score is calculated for the length of the predicted signal peptide. The S-mean score was in SignalP version 2.0 used as the criteria for discrimination of secretory and non-secretory proteins.

The *D-score* is introduced in SignalP version 3.0 and is a simple average of the S-mean and Y-max score. The score shows superior discrimination performance of secretory and non-secretory proteins to that of the S-mean score which was used in SignalP version 1 and 2.

For non-secretory proteins all the scores represented in the SignalP3-NN output should ideally be very low.

The hidden Markov model calculates the probability of whether the submitted sequence contains a signal peptide or not. The eukaryotic HMM model also reports the probability of a signal anchor, previously named uncleaved signal peptides. Furthermore, the cleavage site is assigned by a probability score together with scores for the n-region, h-region, and c-region of the signal peptide, if such one is found.

TargetP output

- One score for each possible location is presented, along with the name and length of the submitted sequence(s).
- C : Chloroplast, i.e. the sequence contains a chloroplast transit peptide, cTP
- M : Mitochondrion, i.e. the sequence contains a mitochondrial targeting peptide, mTP
- S : Secretory pathway, i.e. the sequence contains a signal peptide,
- SP _ : any other location
- * : "don't know". This character appears if cutoff restrictions were demanded and the winning network output score for a sequence was BELOW the requested cutoff for that category. The asterisk shows that no prediction was done by TargetP (although the output scores and RCs are presented also for these sequences).
- Location with the highest score is the most likely one according to TargetP, and the relation between the scores (the reliability class, see below) may be an indication of how certain the prediction is. The reliability class (RC) is a measure of the size of the difference (diff) between the highest (winning) and the second highest output scores.
- **The lower value on the RC, the safer the prediction on that particular sequence. There are 5 reliability classes, defined as follow: RC 1: diff > 0.800 RC 2: 0.800 > diff > 0.600 RC 3: 0.600 > diff > 0.400 RC 4: 0.400 > diff > 0.200 RC 5: 0.200 > diff**
- If cleavage site prediction is opted for, the predicted length of the presequence (if any was predicted) appears in the rightmost column. The actual cleavage site prediction is performed by SignalP for SPs, and by ChloroP for cTPs. The mTP cleavage site prediction, however, is a TargetP-unique feature. The cutoffs for each of the categories are shown. Default is no cutoffs, but that can be changed on the submission page.

TMHMM output

TMHMM statistics:

- Length: the length of the protein sequence.
- Number of predicted TMHs: The number of predicted transmembrane helices.
- Exp number of AAs in TMHs: The expected number of amino acids intransmembrane helices. If this number is larger than 18 it is very likely to be a transmembrane protein (OR have a signal peptide).
- Exp number, first 60 AAs: The expected number of amino acids in transmembrane helices in the first 60 amino acids of the protein. If this number more than a few, you should be warned that a predicted transmembrane helix in the N-term could be a signal peptide.
- Total prob of N-in: The total probability that the N-term is on the cytoplasmic side of the membrane.
- POSSIBLE N-term signal sequence: a warning that is produced when "Exp number, first 60 AAs" is larger than 10.