Functional Annotation

May 23, 2007

Rama Maiti

Functional Annotation Overview

- What is annotation
- Steps we take to annotate eukaryotic genes
- Software tools we use for functional annotation
- Steps we take to manually annotate or verify an automated annotation

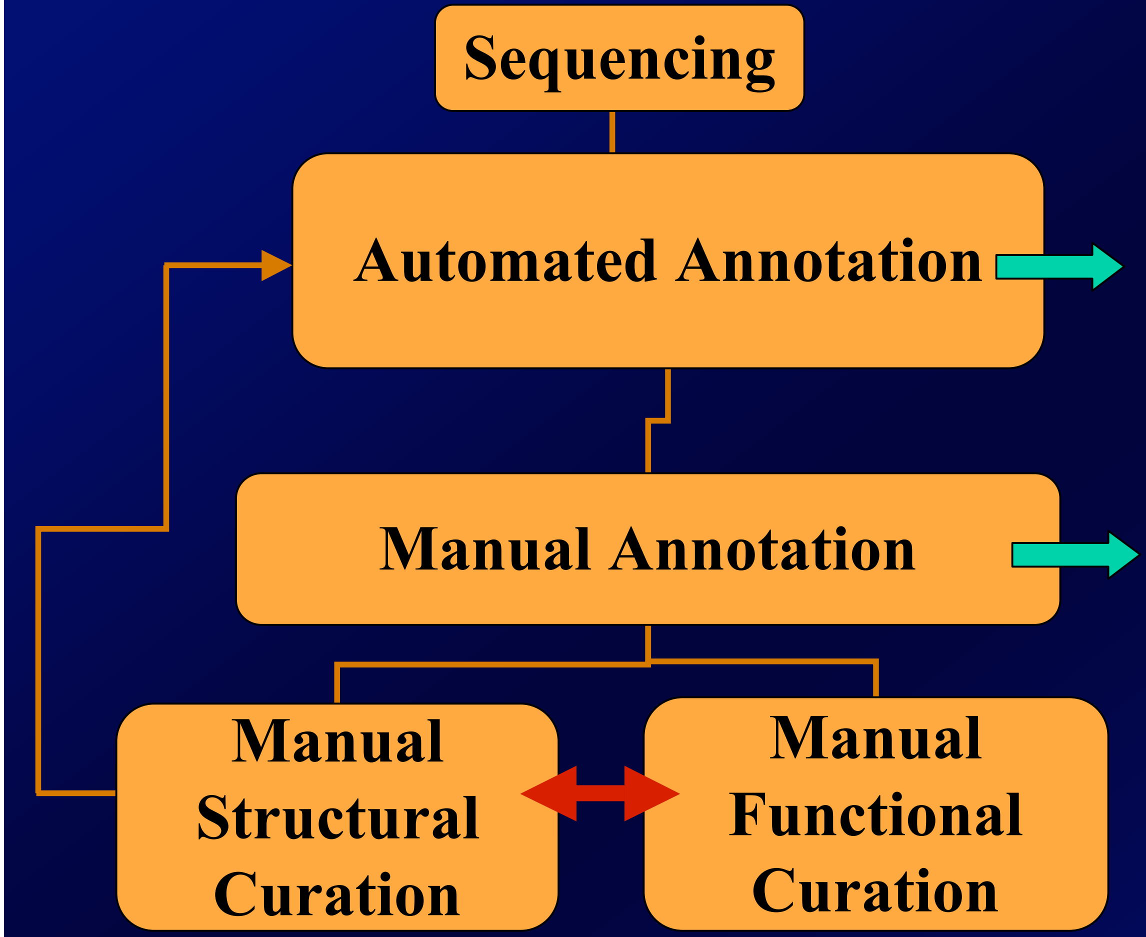
What are the questions?

- How did the gene get its structure and name?
- Does it really have a function assigned to it?
- Where did this information come from?
- Is it accurate? Can you rely on it?

What is Functional Annotation?

- “To annotate” is “to make or furnish critical or explanatory notes or comments”
- For genomics the ‘notes’ are about
  - Names of the gene products
  - Functions of genes within an organism
- Elements of the functional annotation process
  - Validation of the gene structure
  - Literature search, if any is available
  - Homology / domain searches
  - Assignment of function
  - Maintenance of data availability

The Annotation Pipeline



Sequencing

Automated Annotation

Manual Annotation

Manual Structural Curation

Manual Functional Curation

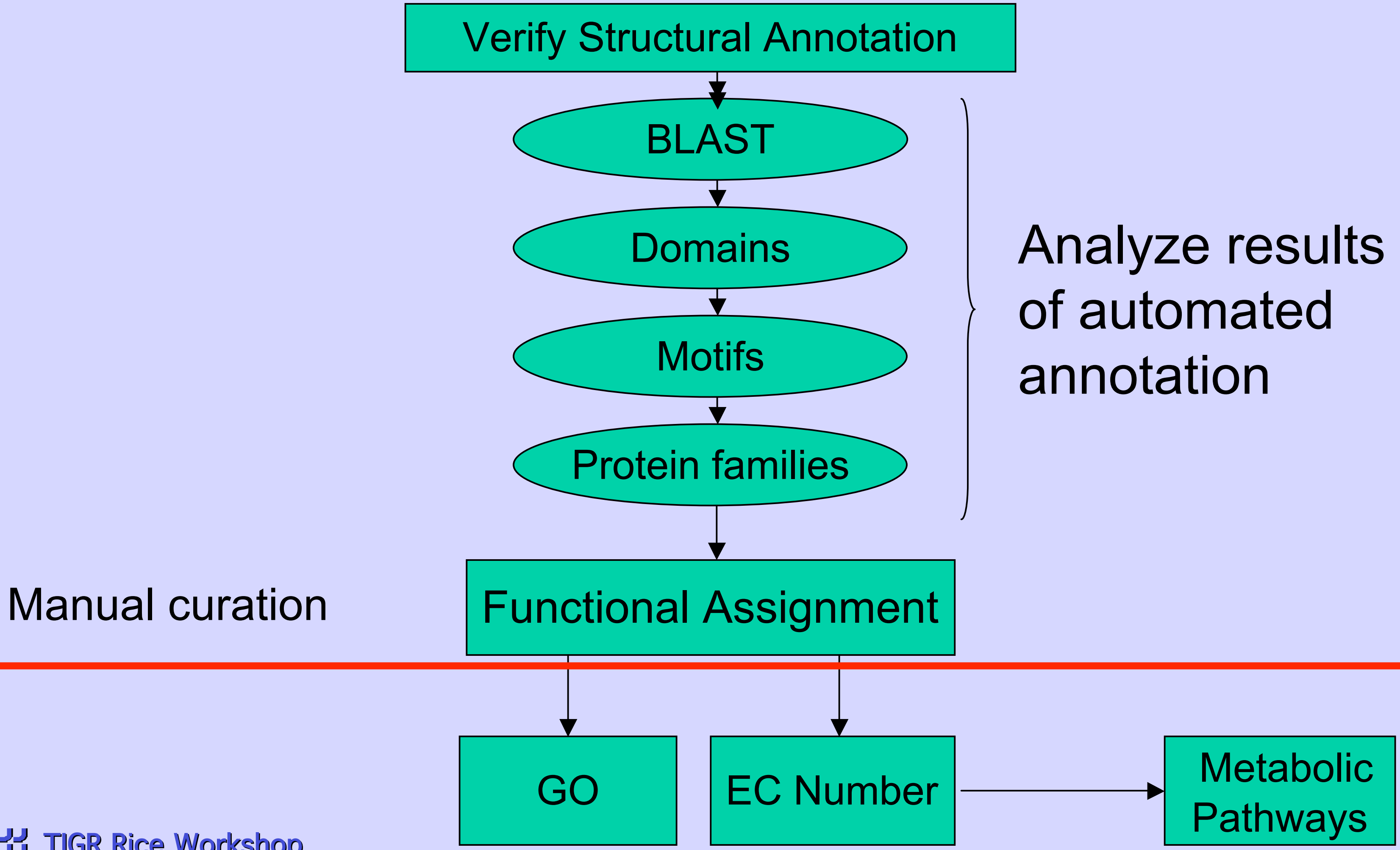
protein searches (nap)
EST, cDNA alignments (gap2)
custom database searches (nap/gap2)
gene prediction algorithms
Blastp searches
HMM searches
SignalP/TargetP/Interpro

Human intervention:
-critical evaluation of
automated assignments

Manual vs. Automated Annotation

- Automated Annotation is complicated by high volumes of data derived from different methods at different centers
- High quality annotation requires manual review and intervention.

Steps in Functional Annotation



Steps in Functional Annotation

- Analyze the gene structure (Annotation Station or preferred gene viewer)
- Name the gene product (Manatee)
  - requires analysis of the gene product
  - gene product name is primarily homology based on different bioinformatics tools
- Assign Gene Ontology terms
  - Process
  - Function
  - Component

Homology Searching

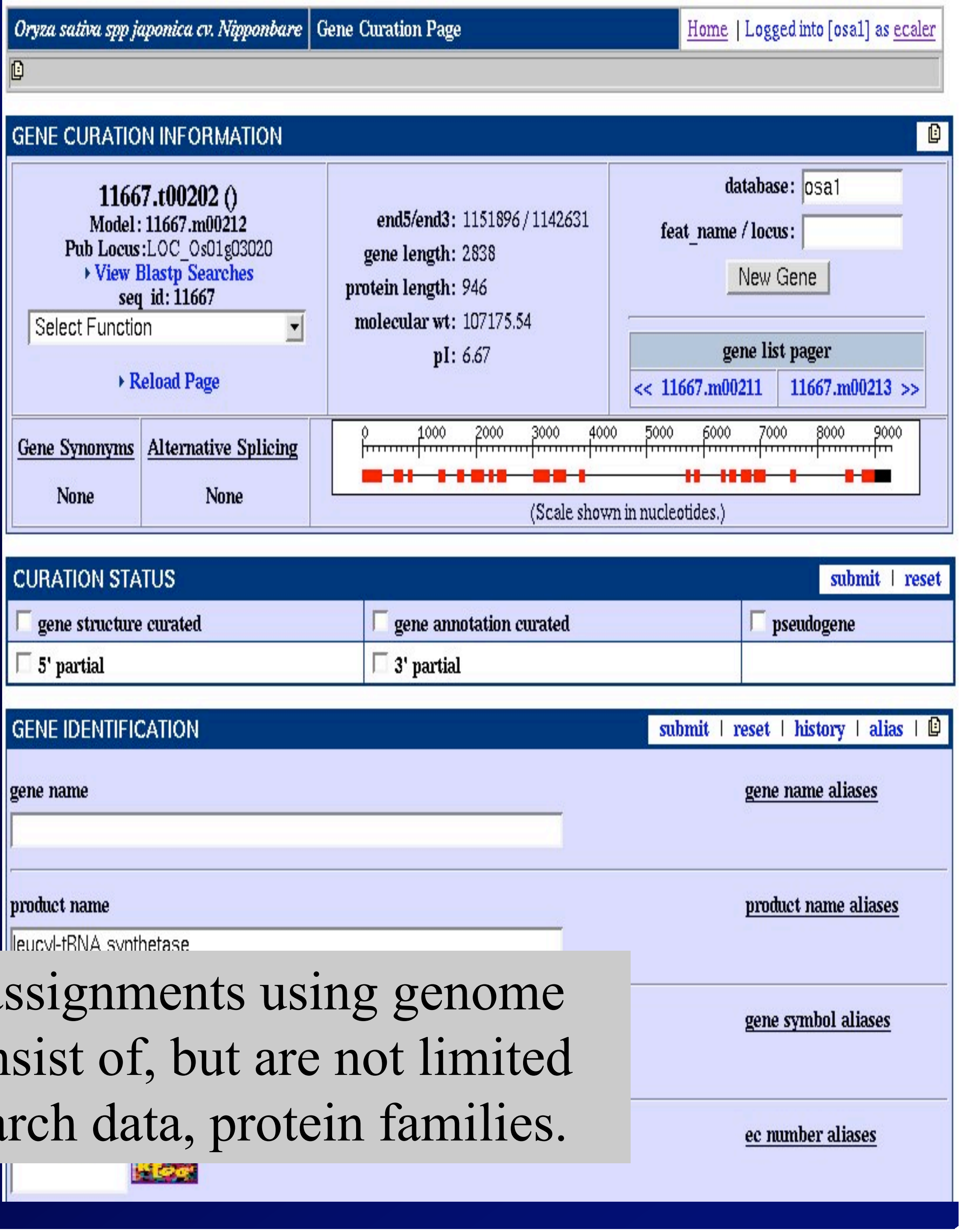
(Tools that are available to characterize a sequence)

- **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/
- **Prosite** (profiles & families) http://ca.expasy.org/tools/scanprosite/
- **Interpro** (families) http://www.ebi.ac.uk/InterProScan/
- **TmHMM** (transmembrane domain) http://www.cbs.dtu.dk/services/TMHMM/
- **Swiss-Prot** **http://au.expasy.org/sprot/**
- **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**
  - **TIGR Paralogous Families** (not yet available outside of TIGR)
  - **TribeMCL** http://www.ebi.ac.uk/research/cgg/tribe/

Manatee

- Manatee is a web-based gene evaluation and genome annotation tool.
- Manatee displays the current annotation for prokaryotic and eukaryotic genomes.
- Manatee is an open source software available at:

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



Oryza sativa spp japonica cv. Nipponbare | Gene Curation Page | Home | Logged into [osa1] as ecaler

GENE CURATION INFORMATION

11667.t00202 ()
Model: 11667.m00212
Pub Locus:LOC_Os01g03020
‣ View Blastp Searches
seq_id: 11667
Select Function
‣ Reload Page

end5/end3: 1151896 / 1142631
gene length: 2838
protein length: 946
molecular wt: 107175.54
pI: 6.67

database: osa1
feat_name / locus:
New Gene

gene list pager
<< 11667.m00211 | 11667.m00213 >>

Gene Synonyms	Alternative Splicing
None	None

(Scale shown in nucleotides.)

CURATION STATUS — submit | reset

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

GENE IDENTIFICATION — submit | reset | history | alias

gene name — gene name aliases

product name: leucyl-tRNA synthetase — product name aliases

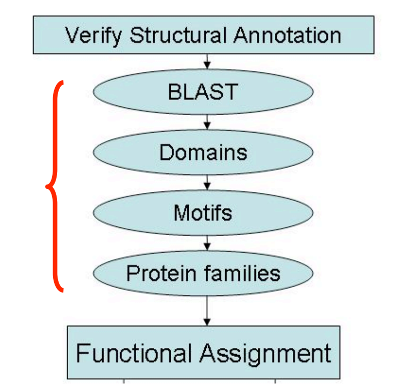
gene symbol aliases

ec number aliases

make high quality functional assignments using genome analyses tools. These tools consist of, but are not limited to GO classifications, blast search data, protein families.

Verify evidence from automated annotation

- BLAST matches
- HMM
- Prosite, Interpro classifications
- Motifs
- Signal Sequence
- Target Sequence
- Transmembrane domain
- Protein families



Functional annotation

Examine the gene structure

does it make sense with respect to the alignments?
do you need to re-curate the gene structure?

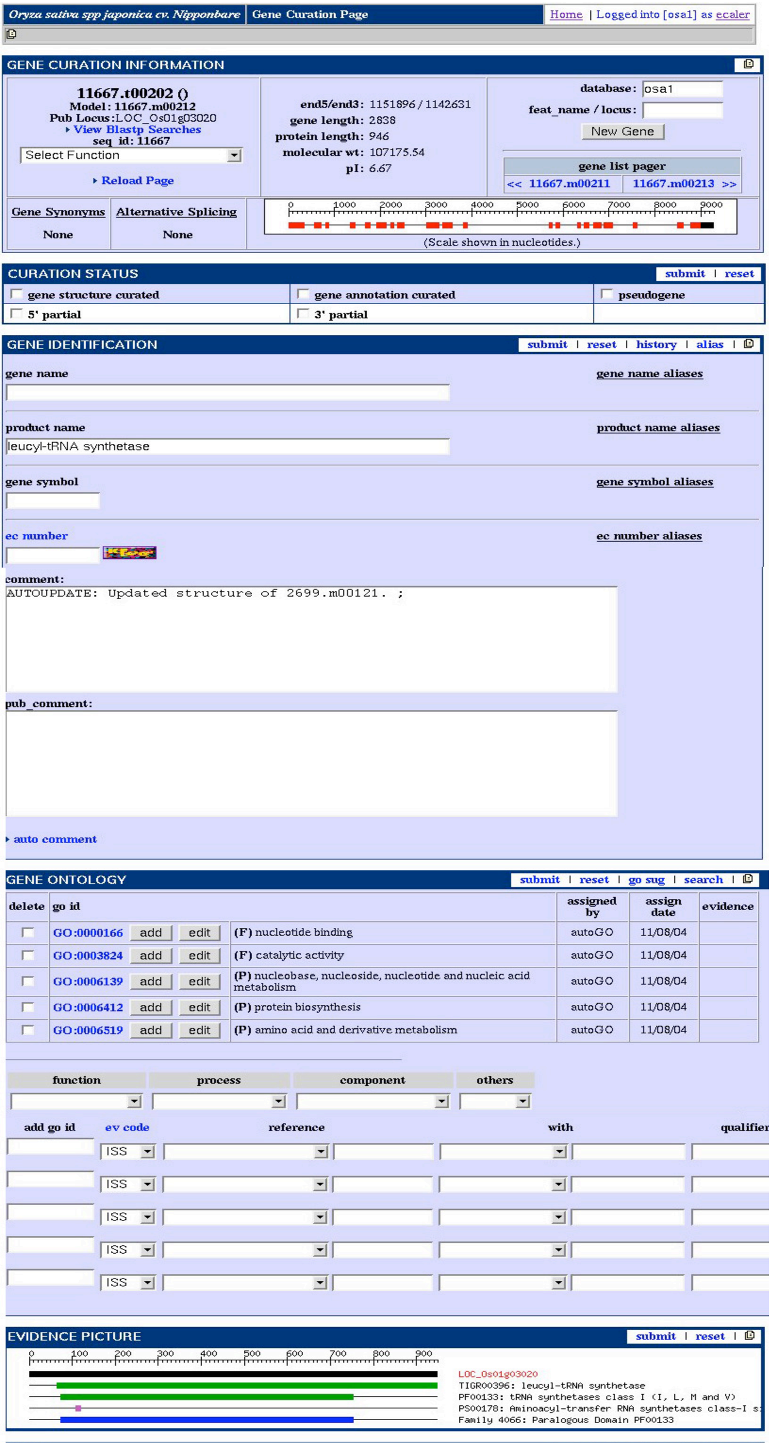
Name the gene product

Determine whether it is published,
Fully characterized? Give it the Swiss-Prot name.
Sequenced but not characterized? Look at the evidence.

Add comments to comment field

explain reasoning for others
add personal communication information
make comments about function or process

In many occasions after analyzing our data and make a decision about a gene function, we may need to go back and re-examine the gene structure.



Oryza sativa spp japonica cv. Nipponbare | Gene Curation Page

GENE CURATION INFORMATION

11667.t00202 ()
Model: 11667.m00212
Pub Locus: LOC_Os01g03020
seq id: 11667

end5/end3: 1151896 / 1142631
gene length: 2838
protein length: 946
molecular wt: 107175.54
pI: 6.67

database: osa1
feat_name / locus:

gene list pager
<< 11667.m00211 | 11667.m00213 >>

Gene Synonyms	Alternative Splicing
None	None

(Scale shown in nucleotides.)

CURATION STATUS

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

GENE IDENTIFICATION

gene name:

gene name aliases

product name: leucyl-tRNA synthetase

product name aliases

gene symbol:

gene symbol aliases

ec number:

ec number aliases

comment:

```
AUTOUPDATE: Updated structure of 2699.m00121. ;
```

pub_comment:

GENE ONTOLOGY

delete	go id				assigned by	assign date	evidence
□	GO:0000166	add	edit	(F) nucleotide binding	autoGO	11/08/04	
□	GO:0003824	add	edit	(F) catalytic activity	autoGO	11/08/04	
□	GO:0006139	add	edit	(P) nucleobase, nucleoside, nucleotide and nucleic acid metabolism	autoGO	11/08/04	
□	GO:0006412	add	edit	(P) protein biosynthesis	autoGO	11/08/04	
□	GO:0006519	add	edit	(P) amino acid and derivative metabolism	autoGO	11/08/04	

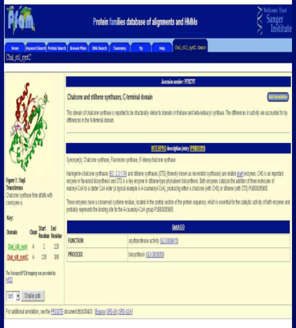
function	process	component	others

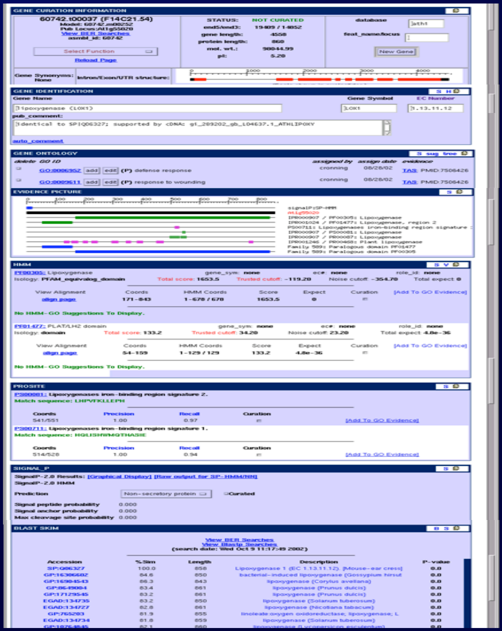
add go id	ev code	reference	with	qualifier
	ISS			
	ISS			
	ISS			
	ISS			
	ISS			

EVIDENCE PICTURE

LOC_Os01g03020
TIGR00396: leucyl-tRNA synthetase
PF00133: tRNA synthetases class I (I, L, M and V)
PS00178: Aminoacyl-transfer RNA synthetases class-I s
Family 4066: Paralogous Domain PF00133

Use all possible resources…


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



GENE CURATION INFORMATION

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

Select Function
Reload Page

STATUS: NOT CURATED
end5/end3: 19409 / 14852
gene length: 4558
protein length: 860
mol. wt.: 98044.99
pI: 5.20

database: ath1
feat_name/locus
New Gene

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

0
100
200
300
400
500
600
700
800
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]

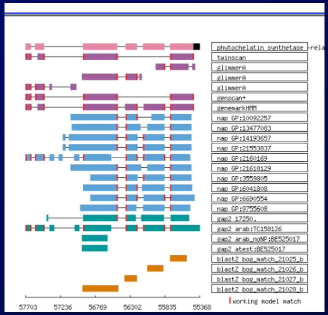
SIGNAL_P

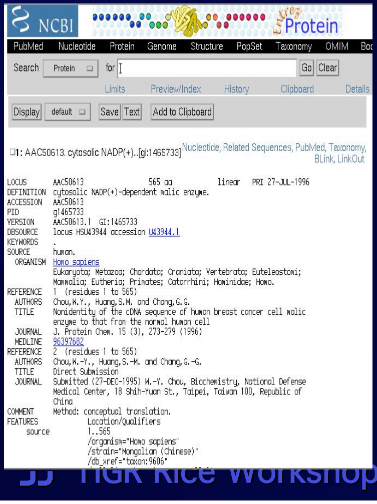
SignalP-2.0 Results: [Graphical Display] [Raw output for SP-HMM/NN]
SignalP-2.0 HMM
Prediction: Non-secretory protein □Curated
Signal peptide probability 0.000
Signal anchor probability 0.000
Max cleavage site probability 0.000

BLAST SKIM

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

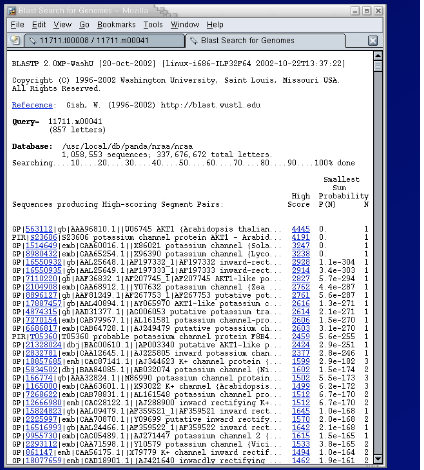
Accession	%Sim	Length	Description	P-value
SP:Q06327	100.0	858	Lipoxygenase 1 (EC 1.13.11.12). [Mouse-ear cress]	0.0
GP:16306602	84.6	850	bacterial-induced lipoxygenase (Gossypium hirsut	0.0
GP:16904543	86.3	843	lipoxygenase (Corylus avellana)	0.0
GP:8649004	83.4	861	lipoxygenase (Prunus dulcis)	0.0
GP:17129545	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:18764845	82.1	860	lipoxygenase (Lycopersicon esculentum)	0.0





□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"
```



```
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
```

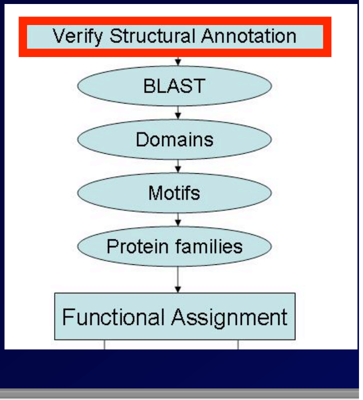
Example:-

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
MLRRLGVRHFRRTPLLFVGGDGSIFERY
TEIDNSNERRINALKGCGMFEDEWIATE
KVHGANFGIYSIEGEKMIRYAKRSGIMP
PNEHFFGYHILIPELQRYITSIREMLCEK
QKKKLHVVLINGELFGGKYDHPSVPKT
RKTVMVAGKPRTISAVQTDSFPQYSPDL
HFYAFDIKYKETEDGDYTTLVYDEAIEL
FQRVPGLLYARAVIRGPMSKVAAFDVE
RFVTTIPPLVGMGNYPLTGNWAEGLVV
KHSRLGMAGFDPKGPTVLKFKCTAFQE
ISTDRAQGPRVDEMRNVRRDSINRAGVQ
LPDLESIVQDPIQLEASKLLLNHVCENRL
KNVLSKIGTEPFEKEEMTPDQLATLLAK
DVLKDFLKDTEPSIVNIPVLIRKDLTRYV
IFESRRLVCSQWKDILKRQSPDFSE*
```

Verify the gene structure



Verify Structural Annotation

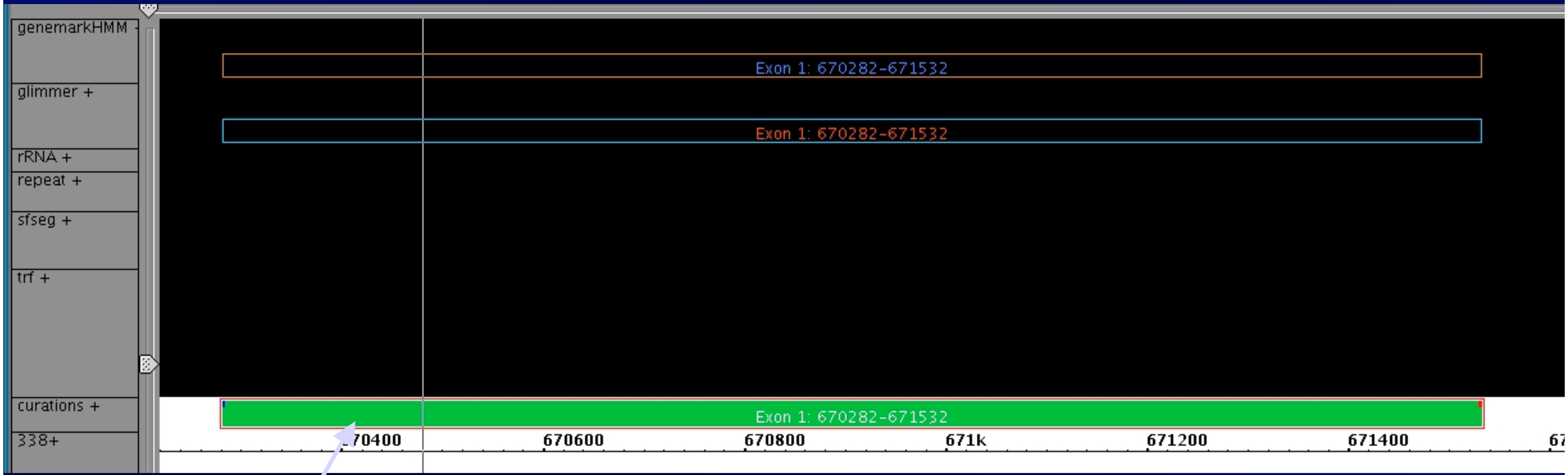
BLAST

Domains

Motifs

Protein families

Functional Assignment



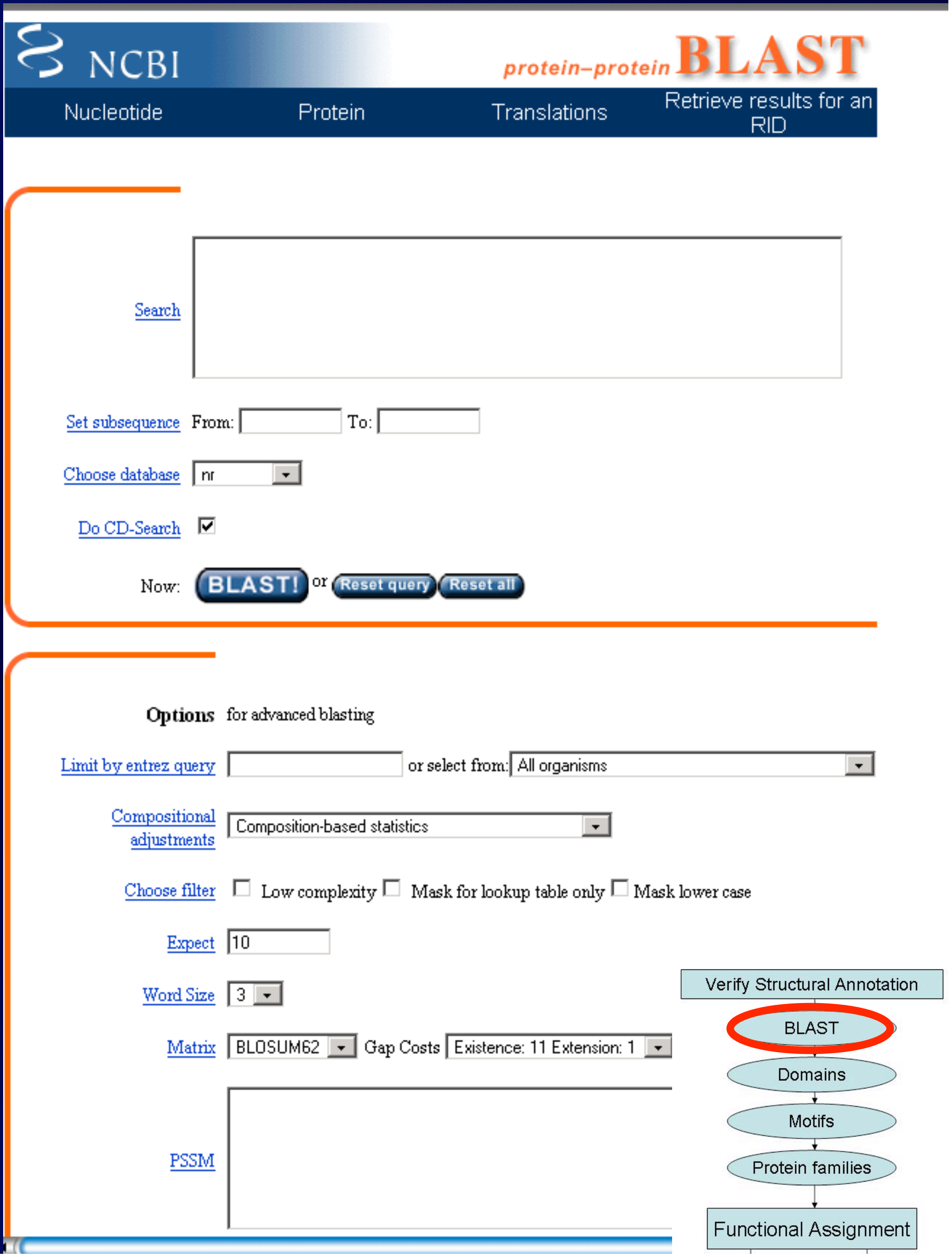
Gene

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

Read → as "translated to"



Search

Set subsequence From: To:

Choose database nr

Do CD-Search ☑

Now: BLAST! or Reset query Reset all

Options for advanced blasting

Limit by entrez query or select from: All organisms

Compositional adjustments Composition-based statistics

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

Expect 10

Word Size 3

Matrix BLOSUM62 Gap Costs Existence: 11 Extension: 1

PSSM

Verify Structural Annotation → BLAST → Domains → Motifs → Protein families → Functional Assignment

BLAST: What makes a good alignment?

It depends on what you are trying to prove!

- minimum of 30% identity, better 40% & up
  - higher for short proteins
  - score is weighted for length

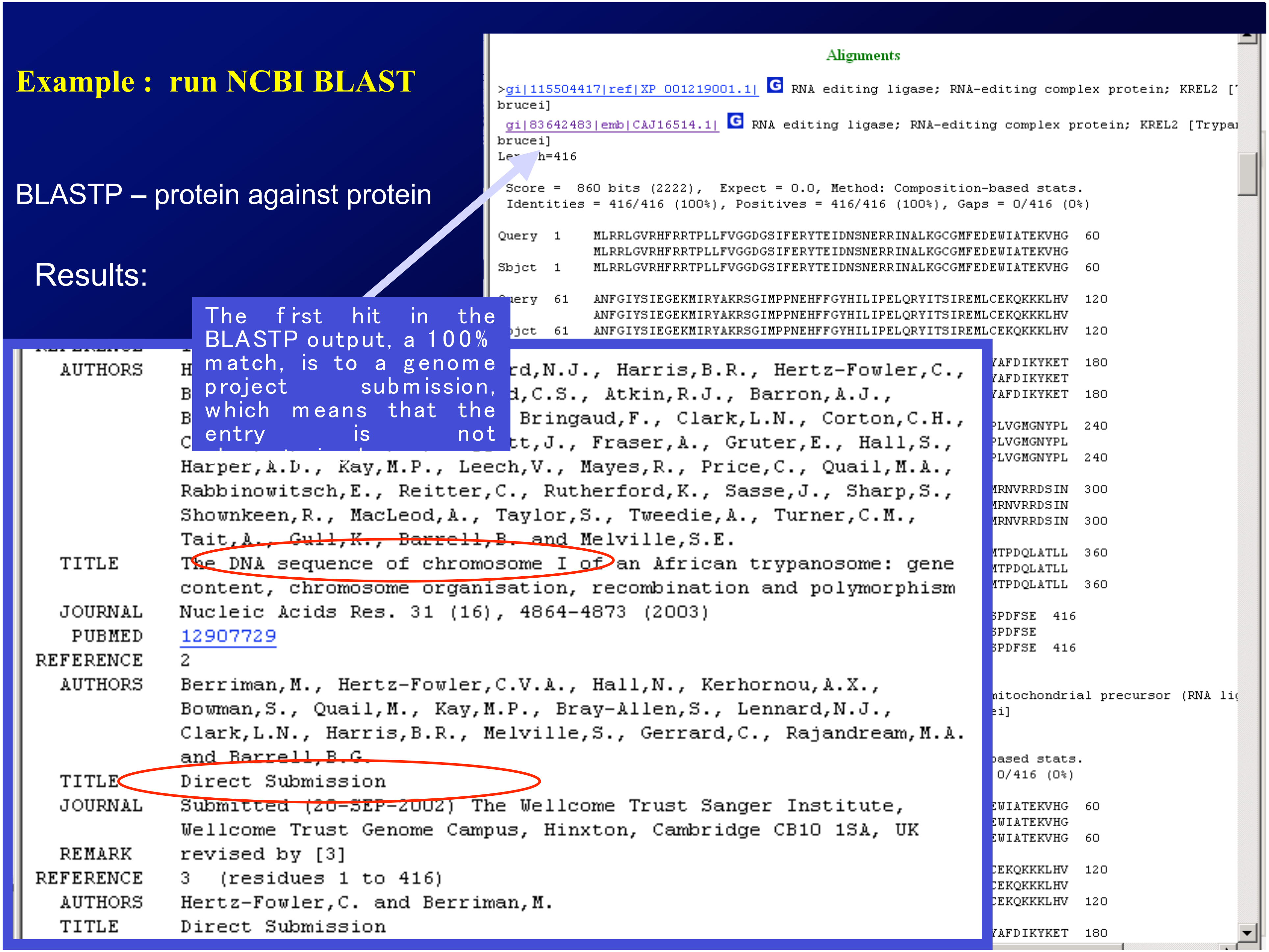
- full length match
  - at least 70% of both proteins

? See explanation of BLAST scores on slide 56.

Example : run NCBI BLAST

BLASTP – protein against protein

Results:

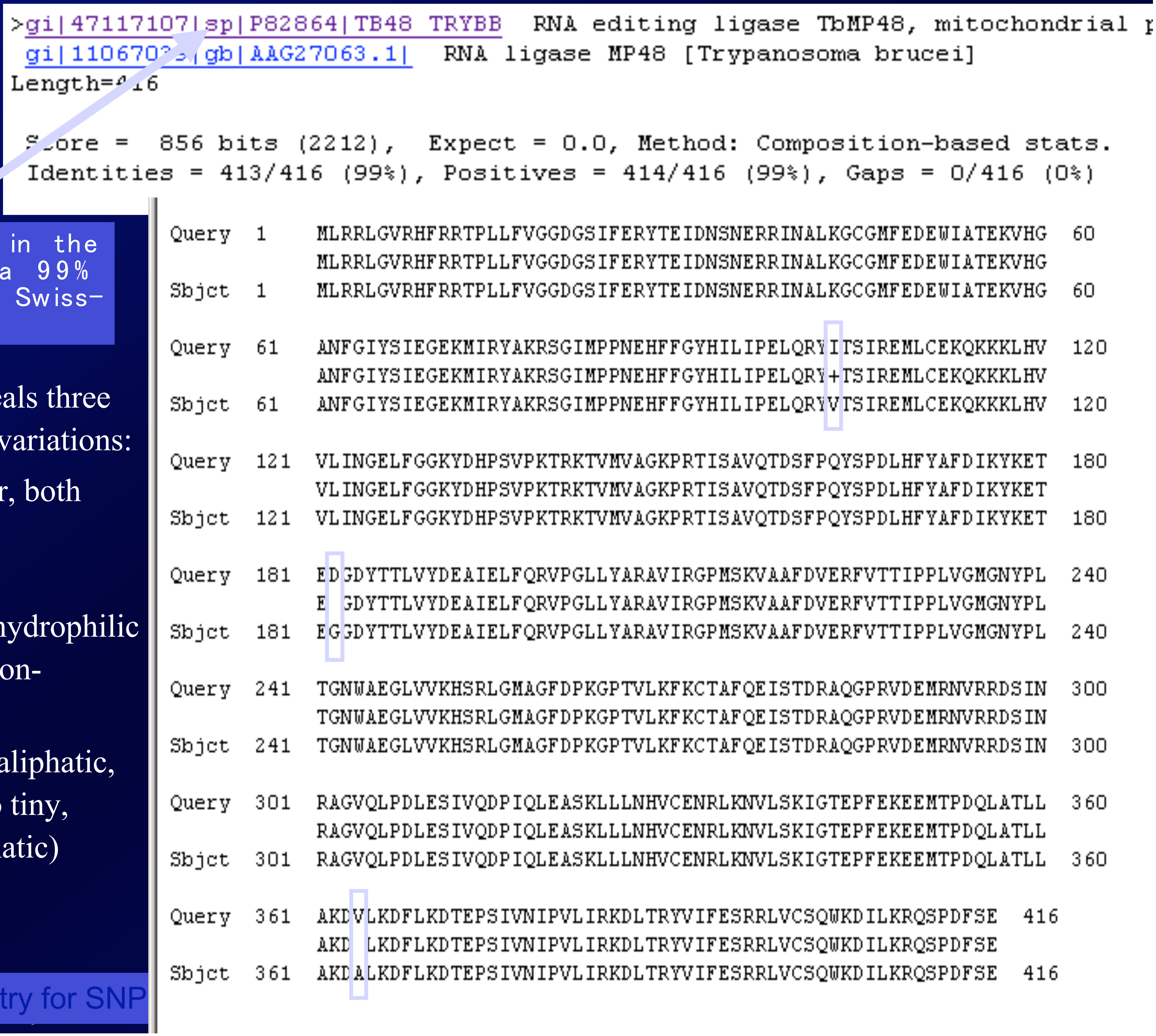


```
>gi|115504417|ref|XP_001219001.1| G RNA editing ligase; RNA-editing complex protein; KREL2 [T
brucei]
 gi|83642483|emb|CAJ16514.1| G RNA editing ligase; RNA-editing complex protein; KREL2 [Trypa
brucei]
Le   h=416

 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
```

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

Example : navigating BLAST output



```
>gi|47117107|sp|P82864|TB48 TRYBB  RNA editing ligase TbMP48, mitochondrial p
 gi|11067033|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 second hit in the BLAST output, a 99% match, is to a Swiss-Prot entry.

The alignment reveals three positions with variations:

I103V (very similar, both hydrophobic) conservative

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

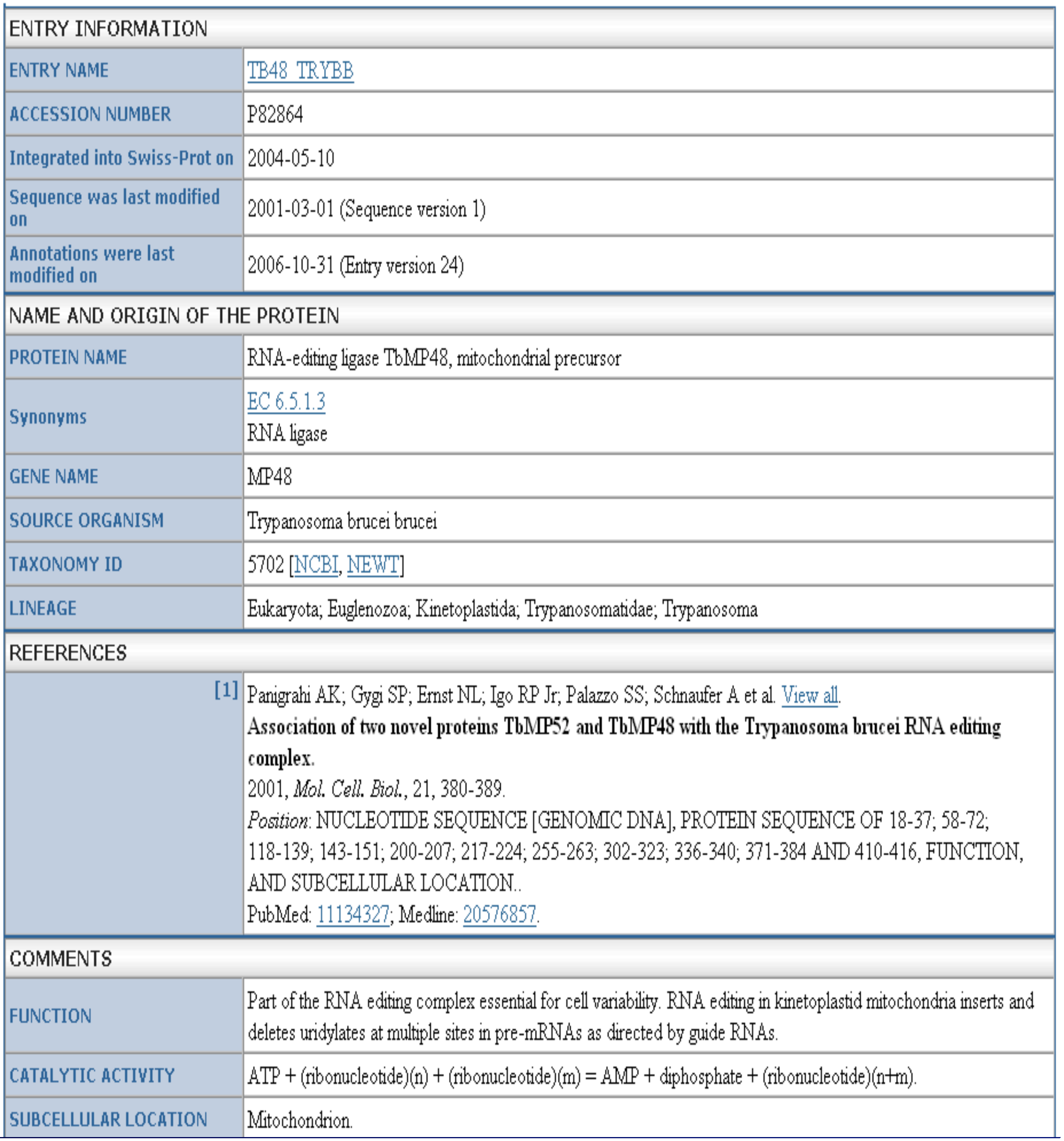
V364A (nonpolar, aliphatic, hydrophobic to tiny, nonpolar, aliphatic) conservative

? See Glossary entry for SNP

Swiss-Prot

Our sequence is 99% identical to the sequence of this Swiss-Prot entry.

Another name for this protein in the literature is ‘REL2.’



ENTRY INFORMATION	
ENTRY NAME	TB48_TRYBB
ACCESSION NUMBER	P82864
Integrated into Swiss-Prot on	2004-05-10
Sequence was last modified on	2001-03-01 (Sequence version 1)
Annotations were last modified on	2006-10-31 (Entry version 24)

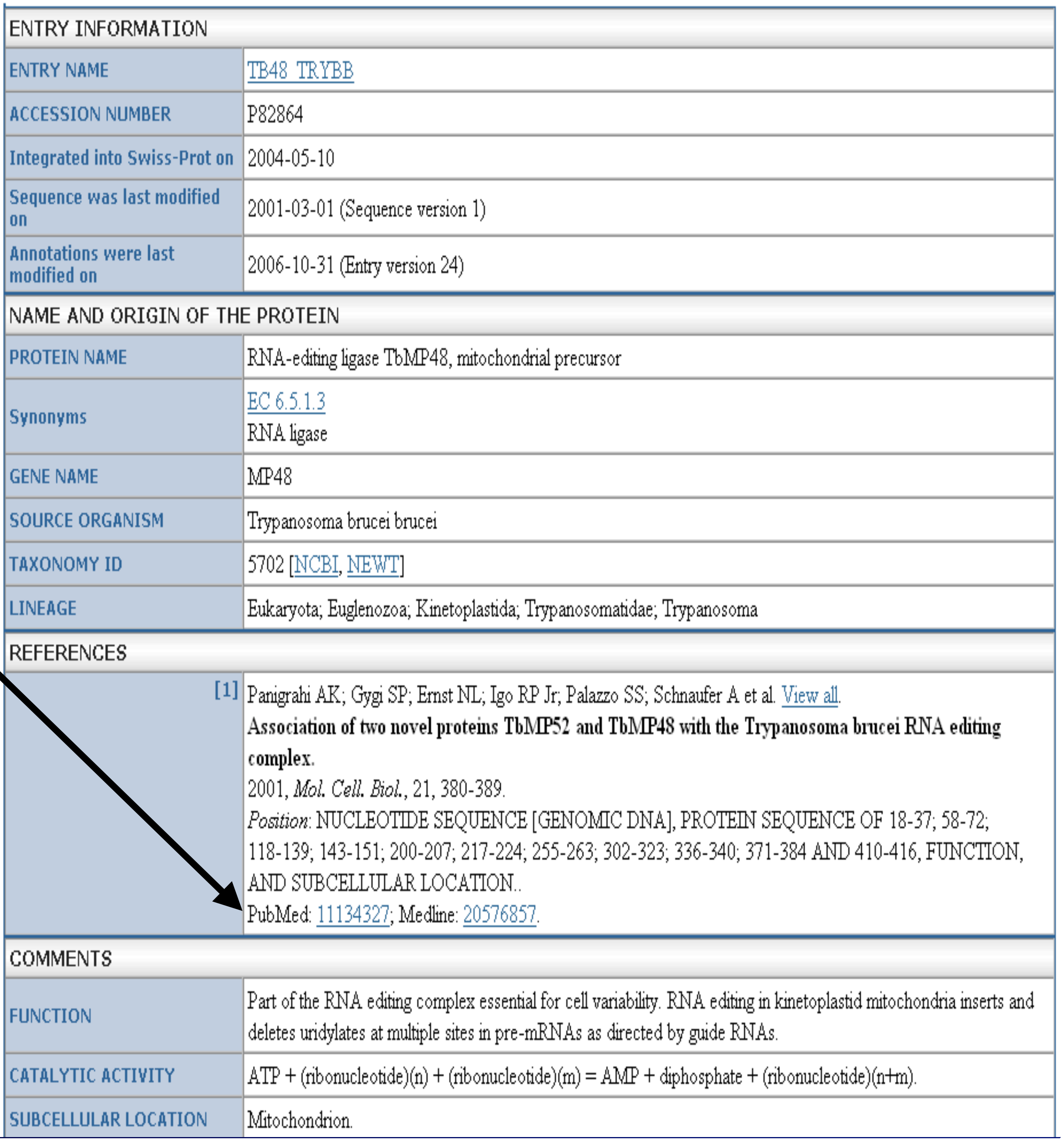
NAME AND ORIGIN OF THE PROTEIN	
PROTEIN NAME	RNA-editing ligase TbMP48, mitochondrial precursor
Synonyms	EC 6.5.1.3 RNA ligase
GENE NAME	MP48
SOURCE ORGANISM	Trypanosoma brucei brucei
TAXONOMY ID	5702 [NCBI, NEWT]
LINEAGE	Eukaryota; Euglenozoa; Kinetoplastida; Trypanosomatidae; Trypanosoma

REFERENCES	
[1]	Panigrahi AK; Gygi SP; Ernst NL; Igo RP Jr; Palazzo SS; Schnaufer A et al. View all. **Association of two novel proteins TbMP52 and TbMP48 with the Trypanosoma brucei RNA editing complex.** 2001, *Mol. Cell. Biol.*, 21, 380-389. *Position:* 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, AND SUBCELLULAR LOCATION.. PubMed: 11134327; Medline: 20576857.

COMMENTS	
FUNCTION	Part of the RNA editing complex essential for cell variability. 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).
SUBCELLULAR LOCATION	Mitochondrion.

Swiss-Prot

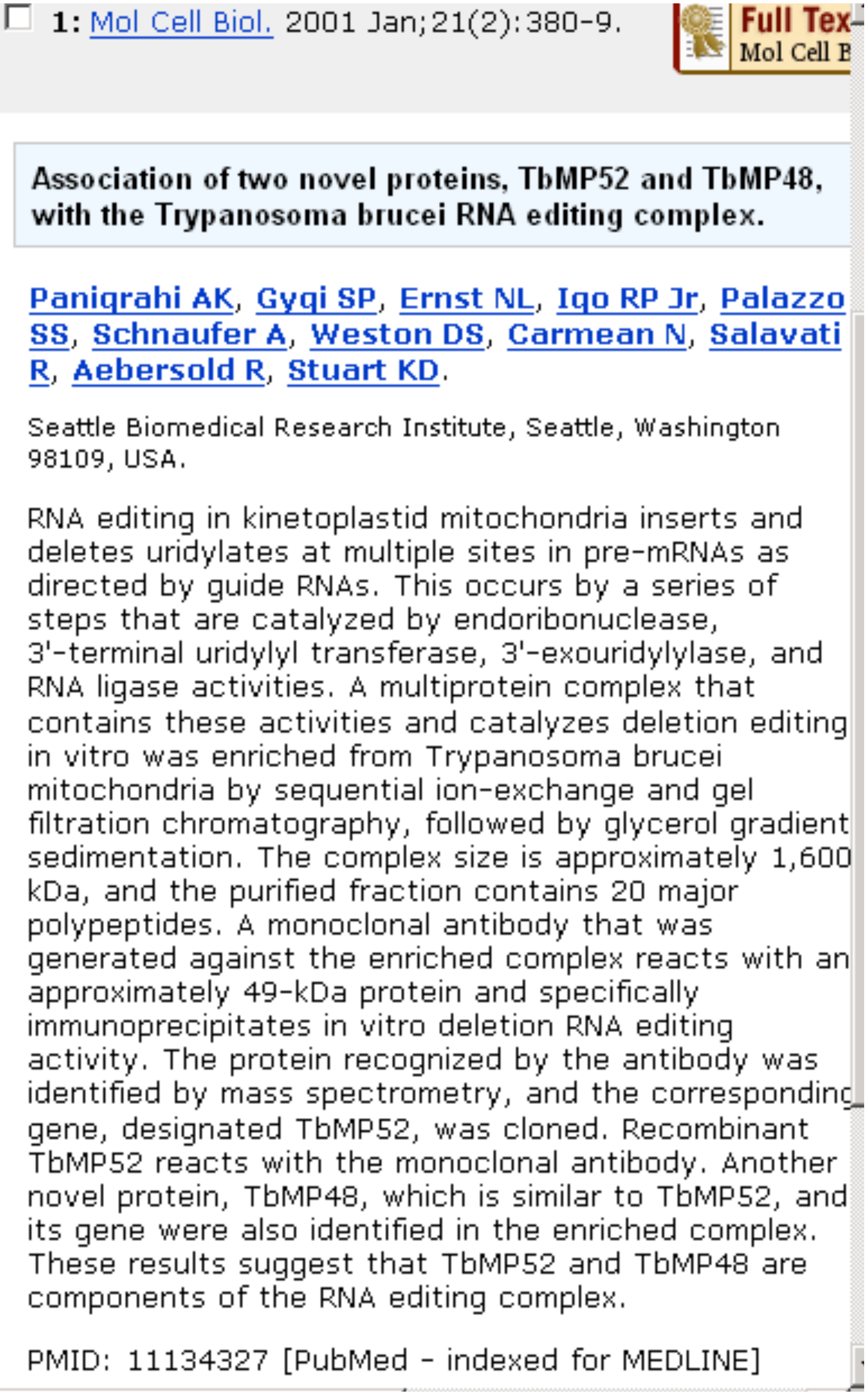
Click on the NCBI hyperlink to look at this publication.



ENTRY INFORMATION	
ENTRY NAME	TB48_TRYBB
ACCESSION NUMBER	P82864
Integrated into Swiss-Prot on	2004-05-10
Sequence was last modified on	2001-03-01 (Sequence version 1)
Annotations were last modified on	2006-10-31 (Entry version 24)
NAME AND ORIGIN OF THE PROTEIN	
PROTEIN NAME	RNA-editing ligase TbMP48, mitochondrial precursor
Synonyms	EC 6.5.1.3 RNA ligase
GENE NAME	MP48
SOURCE ORGANISM	Trypanosoma brucei brucei
TAXONOMY ID	5702 [NCBI, NEWT]
LINEAGE	Eukaryota; Euglenozoa; Kinetoplastida; Trypanosomatidae; Trypanosoma
REFERENCES	
[1]	Panigrahi AK; Gygi SP; Ernst NL; Igo RP Jr; Palazzo SS; Schnaufer A et al. View all. **Association of two novel proteins TbMP52 and TbMP48 with the Trypanosoma brucei RNA editing complex.** 2001, *Mol. Cell. Biol.*, 21, 380-389. *Position*: 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, AND SUBCELLULAR LOCATION.. PubMed: 11134327; Medline: 20576857.
COMMENTS	
FUNCTION	Part of the RNA editing complex essential for cell variability. 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).
SUBCELLULAR LOCATION	Mitochondrion.

Pubmed

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

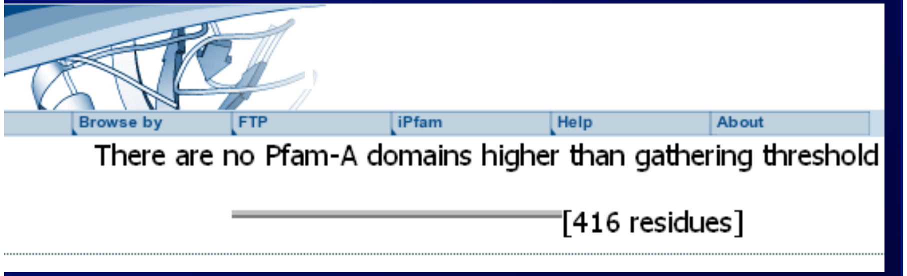
Panigrahi AK, Gygi SP, Ernst NL, Igo RP Jr, Palazzo SS, Schnaufer A, Weston DS, Carmean N, Salavati R, Aebersold R, Stuart KD.

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.

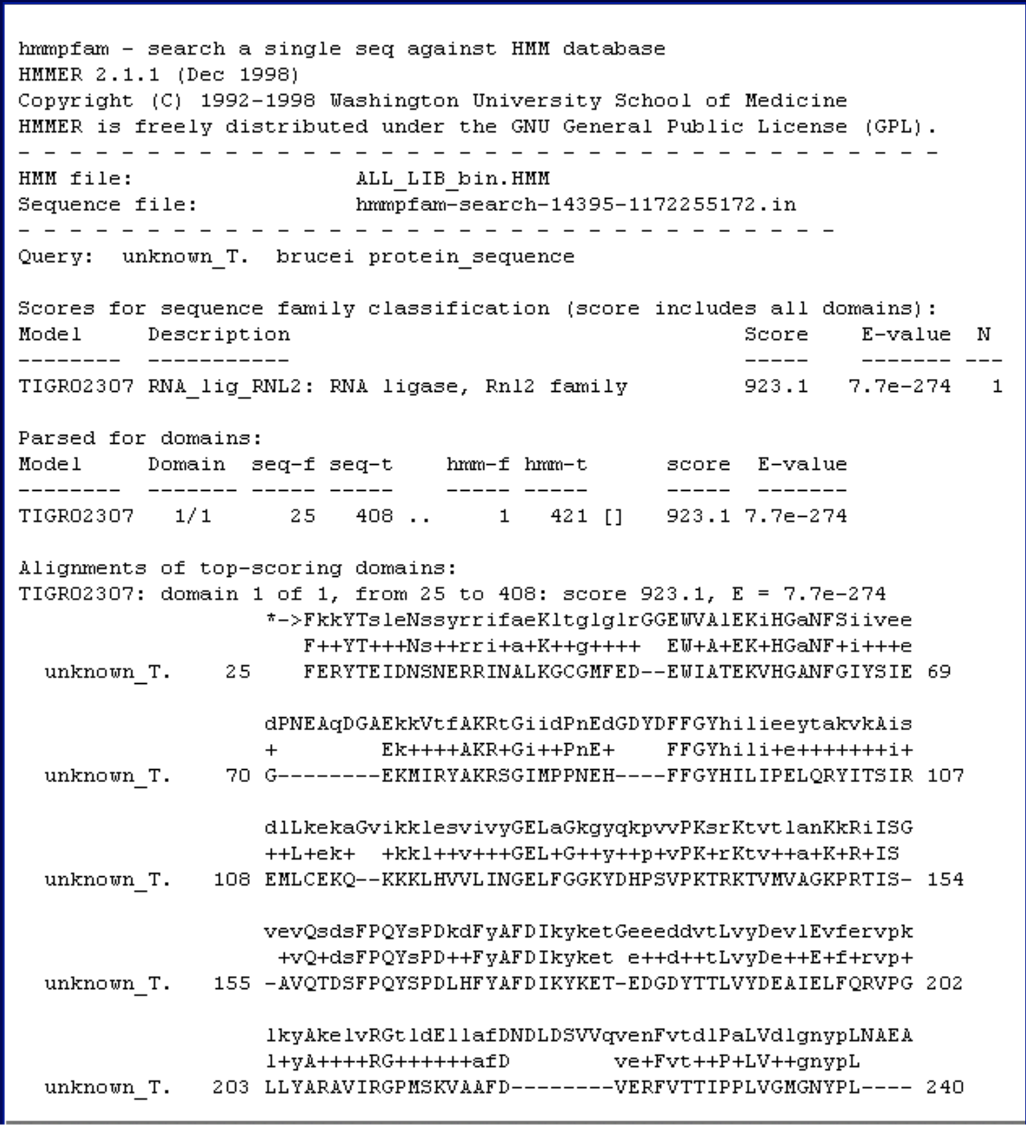
PMID: 11134327 [PubMed - indexed for MEDLINE]

Domains (HMMs) TIGRFAMs search



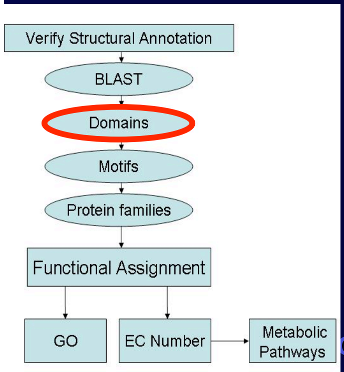
There are no Pfam-A domains higher than gathering threshold

[416 residues]



```
hmmpfam - search a single seq against HMM database
HMMER 2.1.1 (Dec 1998)
Copyright (C) 1992-1998 Washington University School of Medicine
HMMER is freely distributed under the GNU General Public License (GPL).
- - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
HMM file:                 ALL_LIB_bin.HMM
Sequence file:            hmmpfam-search-14395-1172255172.in
- - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
Query:  unknown_T.  brucei protein_sequence

Scores for sequence family classification (score includes all domains):
Model     Description                                   Score    E-value  N
--------  -----------                                   -----    ------- ---
TIGR02307 RNA_lig_RNL2: RNA ligase, Rnl2 family         923.1   7.7e-274   1

Parsed for domains:
Model     Domain  seq-f seq-t    hmm-f hmm-t      score  E-value
--------  ------- ----- -----    ----- -----      -----  -------
TIGR02307   1/1      25   408 ..     1   421 []   923.1 7.7e-274

Alignments of top-scoring domains:
TIGR02307: domain 1 of 1, from 25 to 408: score 923.1, E = 7.7e-274
                   *->FkkYTsleNssyrrifaeKltglglrGGEWVAlEKiHGaNFSiivee
                      F++YT+++Ns++rri+a+K++g++++  EW+A+EK+HGaNF+i+++e
  unknown_T.    25    FERYTEIDNSNERRINALKGCGMFED--EWIATEKVHGANFGIYSIE 69

                   dPNEAqDGAEkkVtfAKRtGiidPnEdGDYDFFGYhilieeytakvkAis
                   +        Ek++++AKR+Gi++PnE+    FFGYhili+e+++++++i+
  unknown_T.    70 G--------EKMIRYAKRSGIMPPNEH----FFGYHILIPELQRYITSIR 107

                   dlLkekaGvikklesvivyGELaGkgyqkpvvPKsrKtvtlanKkRiISG
                   ++L+ek+  +kkl++v+++GEL+G++y++p+vPK+rKtv++a+K+R+IS
  unknown_T.   108 EMLCEKQ--KKKLHVVLINGELFGGKYDHPSVPKTRKTVMVAGKPRTIS- 154

                   vevQsdsFPQYsPDkdFyAFDIkyketGeeeddvtLvyDevlEvfervpk
                    +vQ+dsFPQYsPD++FyAFDIkyket e++d++tLvyDe++E+f+rvp+
  unknown_T.   155 -AVQTDSFPQYSPDLHFYAFDIKYKET-EDGDYTTLVYDEAIELFQRVPG 202

                   lkyAkelvRGtldEllafDNDLDSVVqvenFvtdlPaLVdlgnypLNAEA
                   l+yA++++RG++++++afD        ve+Fvt++P+LV++gnypL
  unknown_T.   203 LLYARAVIRGPMSKVAAFD--------VERFVTTIPPLVGMGNYPL---- 240
```

Total score: 923.1
Trusted cutoff: 100.0
Gathering cutoff: 100.0
Noise cutoff: -165.0

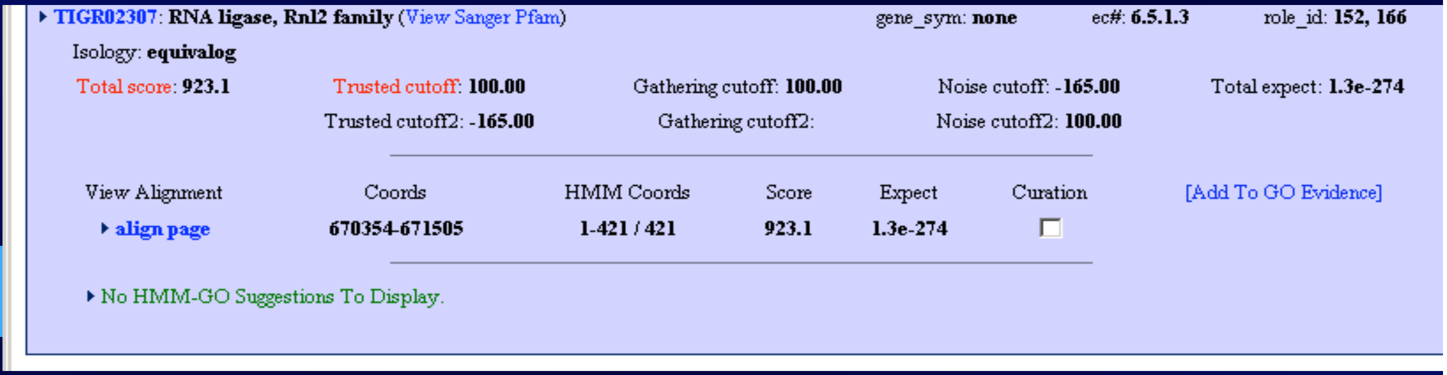


This is a very positive hit to the RNA ligase RNL2 family domain (TIGR02307).

Verify HMM

Total score:	923.1
Trusted cutoff:	100.0
Gathering cutoff:	100.0
Noise cutoff:	-165.0

Score is well above the trusted cutoff.



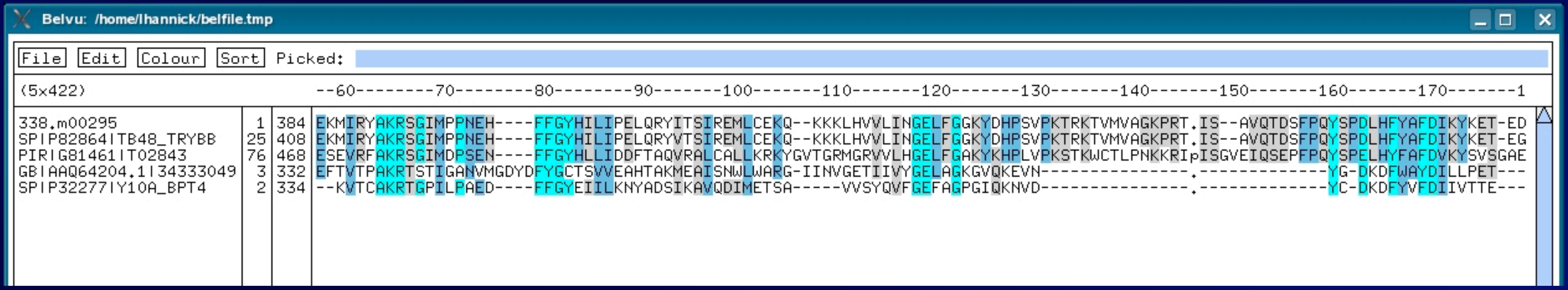
▸ TIGR02307: **RNA ligase, Rnl2 family** (View Sanger Pfam) gene_sym: **none** ec#: **6.5.1.3** role_id: **152, 166**

Isology: **equivalog**

Total score: **923.1**	Trusted cutoff: **100.00**	Gathering cutoff: **100.00**	Noise cutoff: **-165.00**	Total expect: **1.3e-274**
	Trusted cutoff2: **-165.00**	Gathering cutoff2:	Noise cutoff2: **100.00**	

View Alignment	Coords	HMM Coords	Score	Expect	Curation	[Add To GO Evidence]
▸ **align page**	**670354-671505**	**1-421 / 421**	**923.1**	**1.3e-274**	□	

▸ No HMM-GO Suggestions To Display.



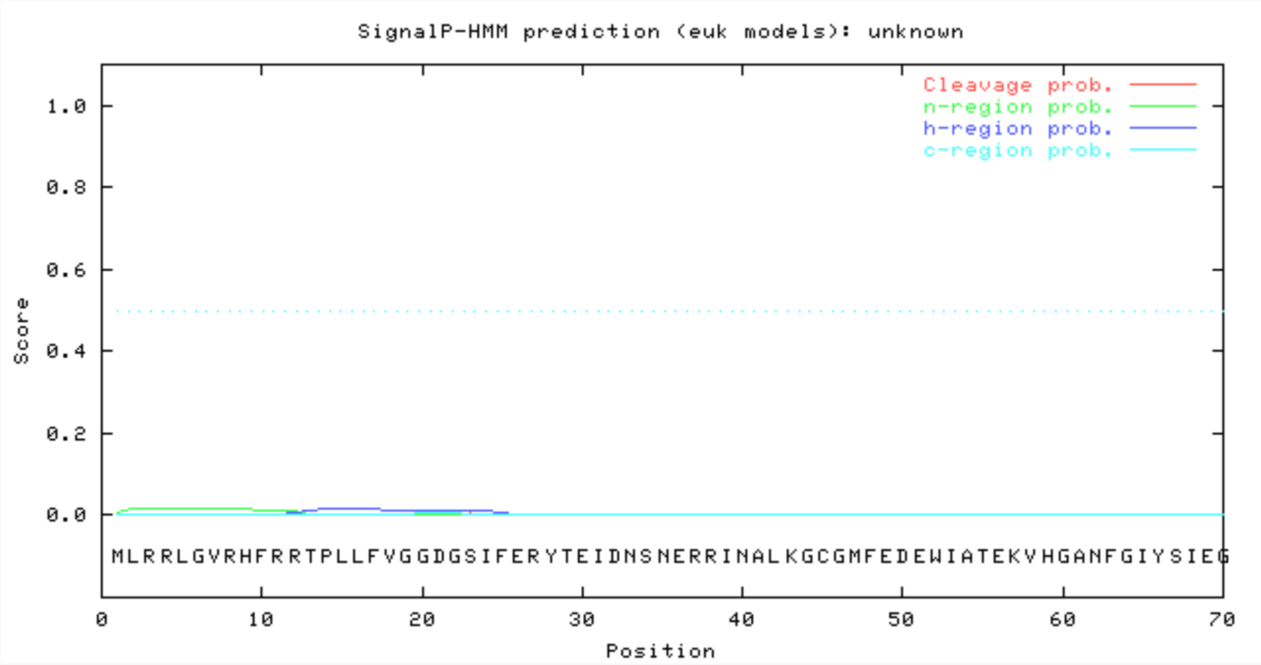
```
(5x422)                                --60--------70--------80--------90-------100-------110-------120-------130-------140-------150-------160-------170-------1
338.m00295                 1 384 EKMIRYAKRSGIMPPNEH----FFGYHILIPELQRYITSIREMLCEKQ--KKKLHVVLINGELFGGKYDHPSVPKTRKTVMVAGKPRT.IS--AVQTDSFPQYSPDLHFYAFDIKYKET-ED
SPIP82864ITB48_TRYBB      25 408 EKMIRYAKRSGIMPPNEH----FFGYHILIPELQRYVTSIREMLCEKQ--KKKLHVVLINGELFGGKYDHPSVPKTRKTVMVAGKPRT.IS--AVQTDSFPQYSPDLHFYAFDIKYKET-EG
PIRIG81461IT02843         76 468 ESEVRFAKRSGIMDPSEN----FFGYHLLIDDFTAQVRALCALLKRKYGVTGRMGRVVLHGELFGAKYKHPLVPKSTKWCTLPNKKRIpISGVEIQSEPFPQYSPELHYFAFDVKYSVSGAE
GBIAAQ64204.1I34333049     3 332 EFTVTPAKRTSTIGANVMGDYDFYGCTSVVEAHTAKMEAISNWLWARG-IINVGETIIVYGELAGKGVQKEVN--------------.-------------YG-DKDFWAYDILLPET---
SPIP32277IY10A_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 TIGRfam family ligate RNA.

Non-secretory protein

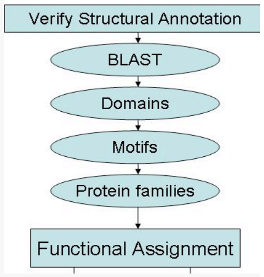
See Glossary entry for Signal Peptide

SignalP-HMM result:



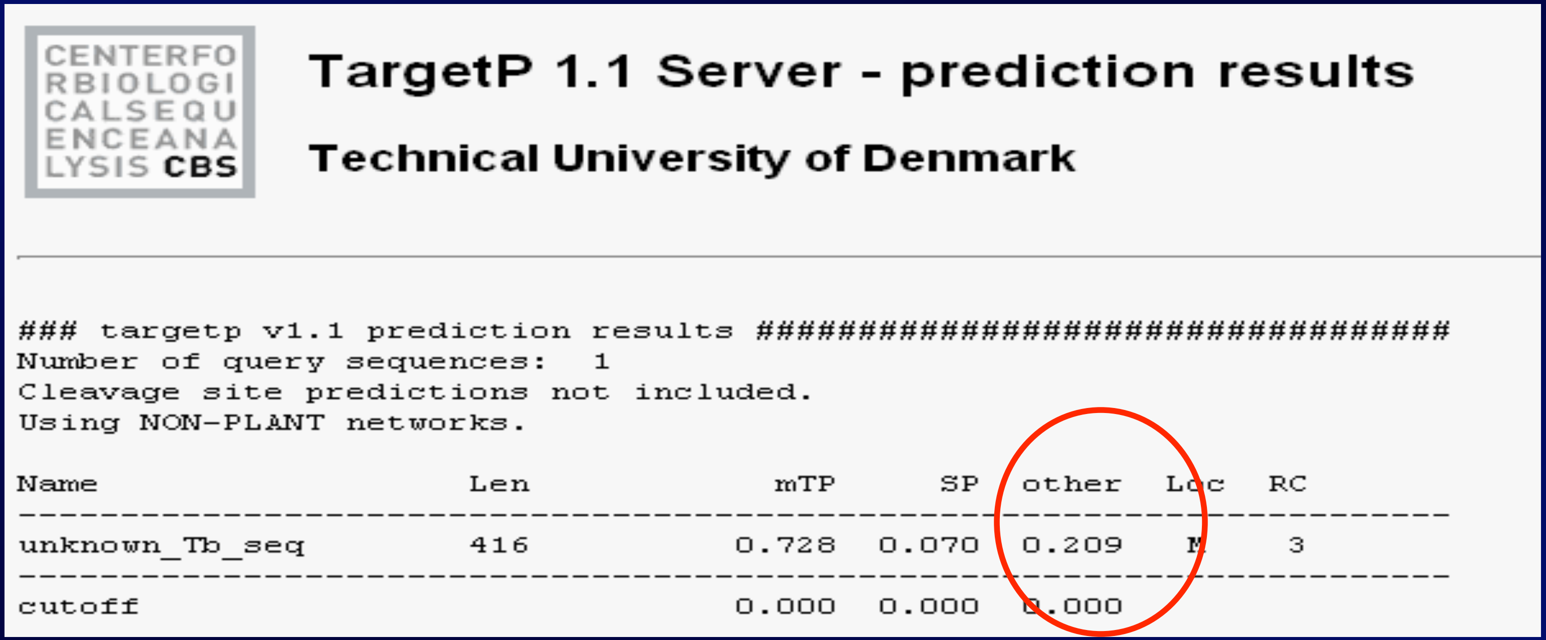
data

```
>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
```



TargetP

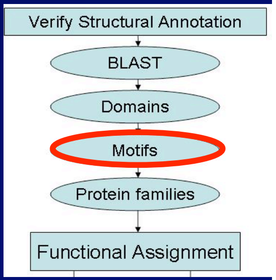
The sequence contains a mitochondrial targeting peptide, mTP.



TargetP 1.1 Server - prediction results

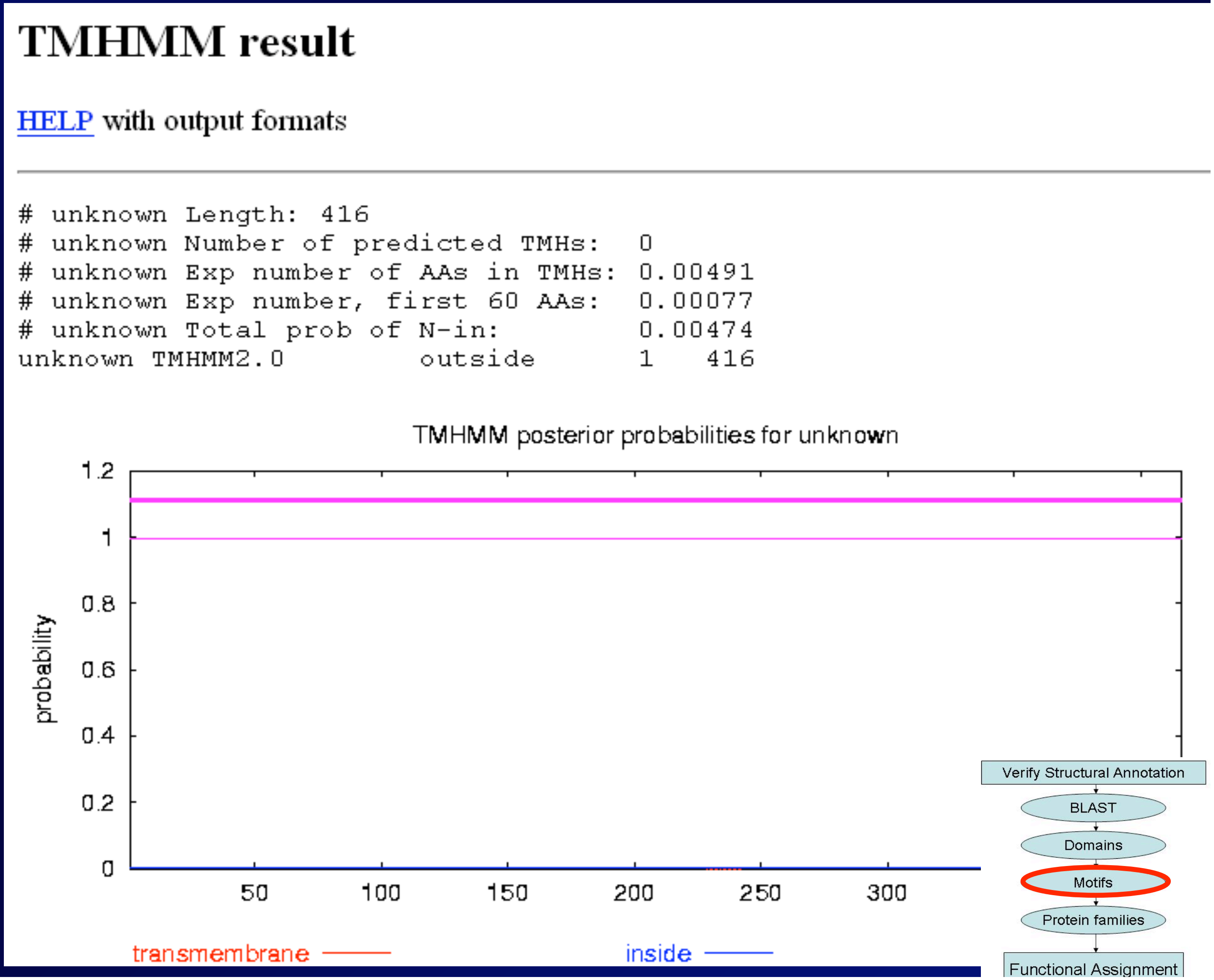
Technical University of Denmark

```
### 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
```



Transmembrane domains

There are no 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
```

Annotation of Example Protein

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.

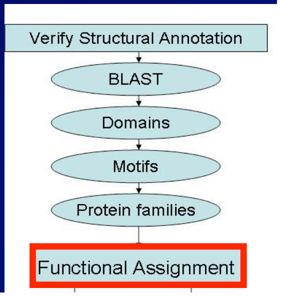
This mitochondrial precursor of an RNA ligase was identified as a member of a multi-protein complex that catalyzes deletion editing in vitro. It was isolated from an enriched sample of Trypanosoma brucei mitochondria by sequential ion-exchange and gel filtration chromatography, followed by glycerol gradient 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 TIGRFAMs domain, with a very strong match. Members of this TIGRfam family ligate (seal breaks in) RNA.

Signal sequence: none

Targeting Sequence: It contains a mitochondrial targeting sequence.

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



Verify Structural Annotation

BLAST

Domains

Motifs

Protein families

Functional Assignment

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 (>30%)
- Domains and motifs
- Binding sites
- Catalytic sites

But :

- 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.

Transitive annotation

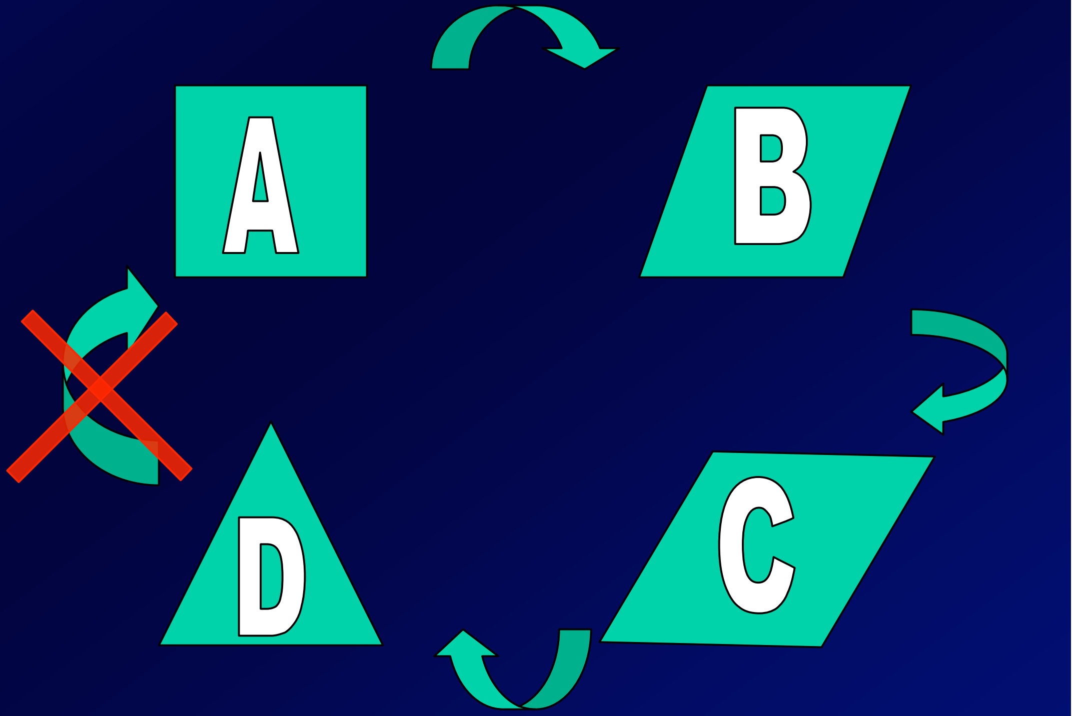
Beware!

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.

Gene Ontology

GO is…

- a method used to structure biological knowledge using a dynamic controlled vocabulary across organisms.
- a database containing a shared vocabulary of descriptive terms for the description of the molecular function, biological process and cellular component of gene products.
- The Gene Ontology Consortium™ is a collaboration among model genome organism databases.

Topics

- Reasons GO has been developed
- Nuts and bolts of GO
- Tools
- Searching GO
- Assigning terms
- GO Slims

The Basics

- GO is a controlled vocabulary
- GO has three aspects, or ontologies:
  - Molecular function
  - Biological process
  - Cellular component
- The 3 aspects refer to genes and gene products

The specificity of GO

There is a limit to how much information can be contained in the name of a protein. For example:

“translation initiation factor 2 subunit”

GO terms assigned to this tell much more:

GO:0003743	(MF)	translation initiation factor activity
GO:0005525	(MF)	GTP binding
GO:0006413	(BP)	translational initiation
GO:0005851	(CC)	eukaryotic translation initiation factor 2B complex

The Gene Ontology is like a dictionary

Each concept has:

- a name
- a definition
- an ID number

term: transcription initiation

id: GO:0006352

definition: Processes involved in the assembly of the RNA polymerase complex at the promoter region of a DNA template resulting in the subsequent synthesis of RNA from that promoter.

GO terms

- A GO term, or ID, is attached to every function, process or component
- There are relationships between them
- Relationships are shown by a graph
  - Directed acyclic graph
  - Sometimes called a “tree”

GO Tools

GO tools are available at the GO Consortium:
http://www.geneontology.org/GO.tools.shtml

Developed and maintained by GO:

AmiGO - Searching through terms and annotations

OBO-Edit - Editing and viewing the DAG

Many others developed independently, for:

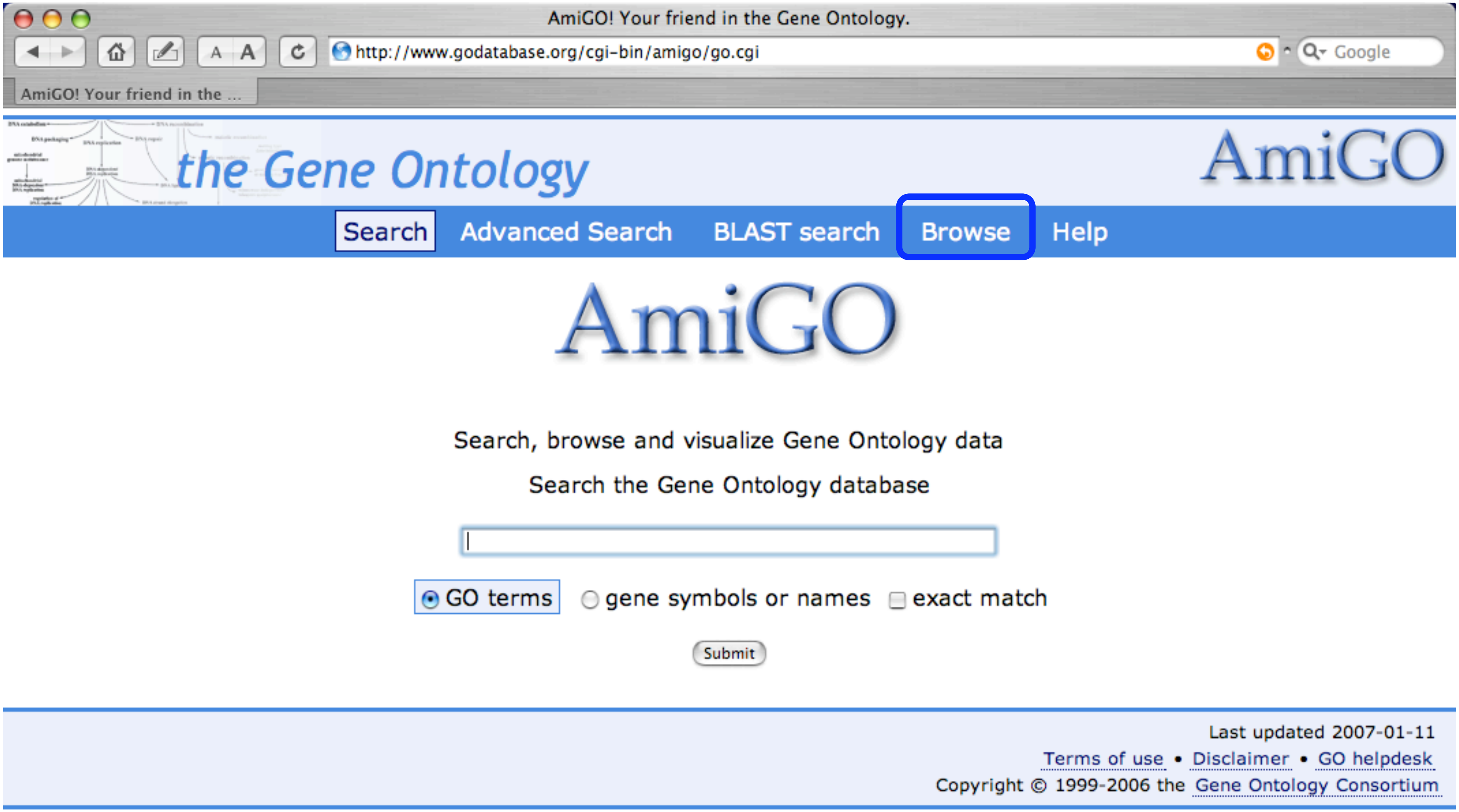
Annotation

Gene expression/microarray data

GO Slims

AmiGO

The GO Browser



http://www.godatabase.org/cgi-bin/amigo/go.cgi

Search, browse and visualize Gene Ontology data

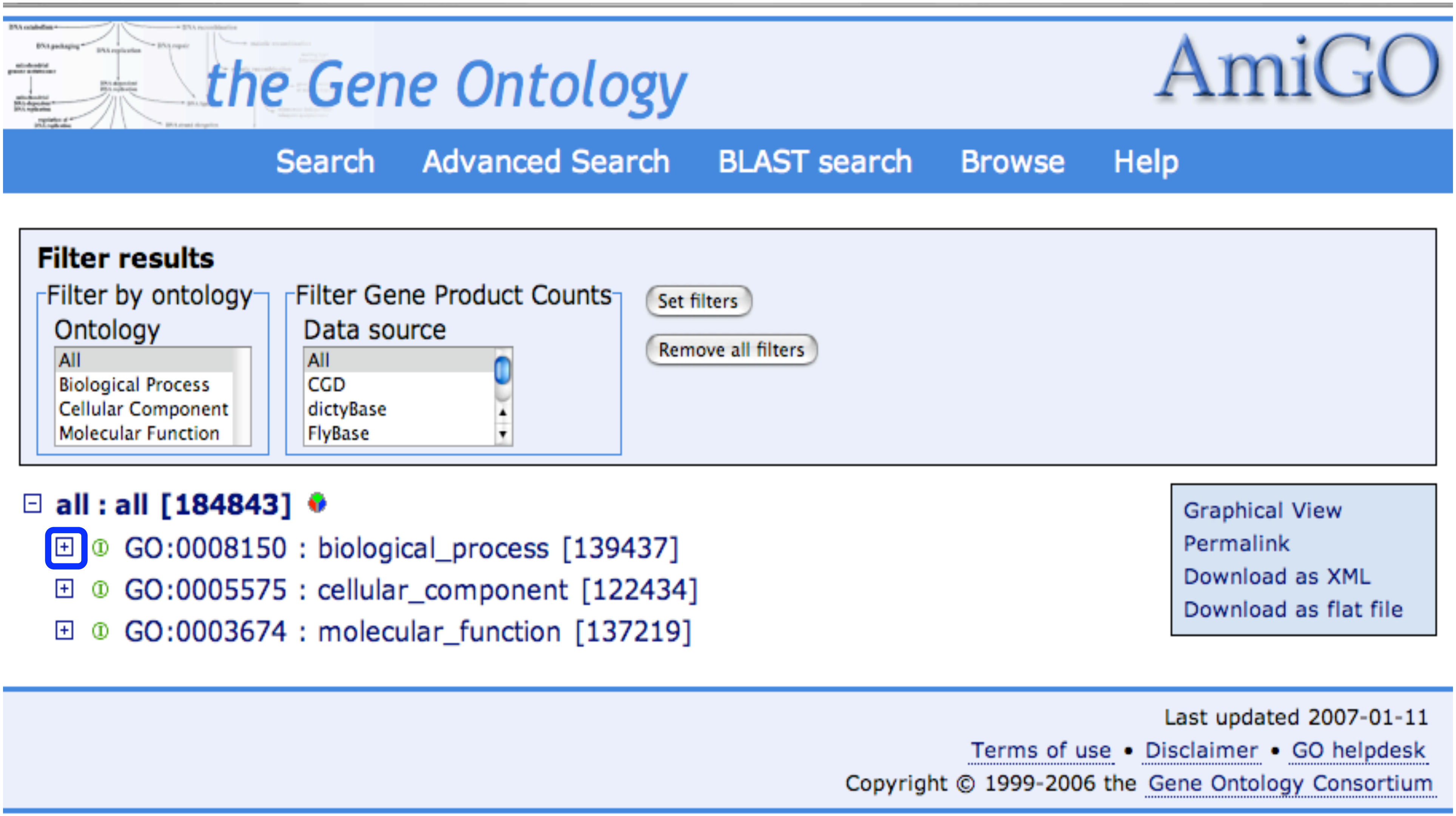
Search the Gene Ontology database

● GO terms ○ gene symbols or names □ exact match

Last updated 2007-01-11

Terms of use • Disclaimer • GO helpdesk

Copyright © 1999-2006 the Gene Ontology Consortium



Filter results

Filter by ontology

Ontology

- All
- Biological Process
- Cellular Component
- Molecular Function

Filter Gene Product Counts

Data source

- All
- CGD
- dictyBase
- FlyBase

⊟ **all : all [184843]**

⊞ GO:0008150 : biological_process [139437]

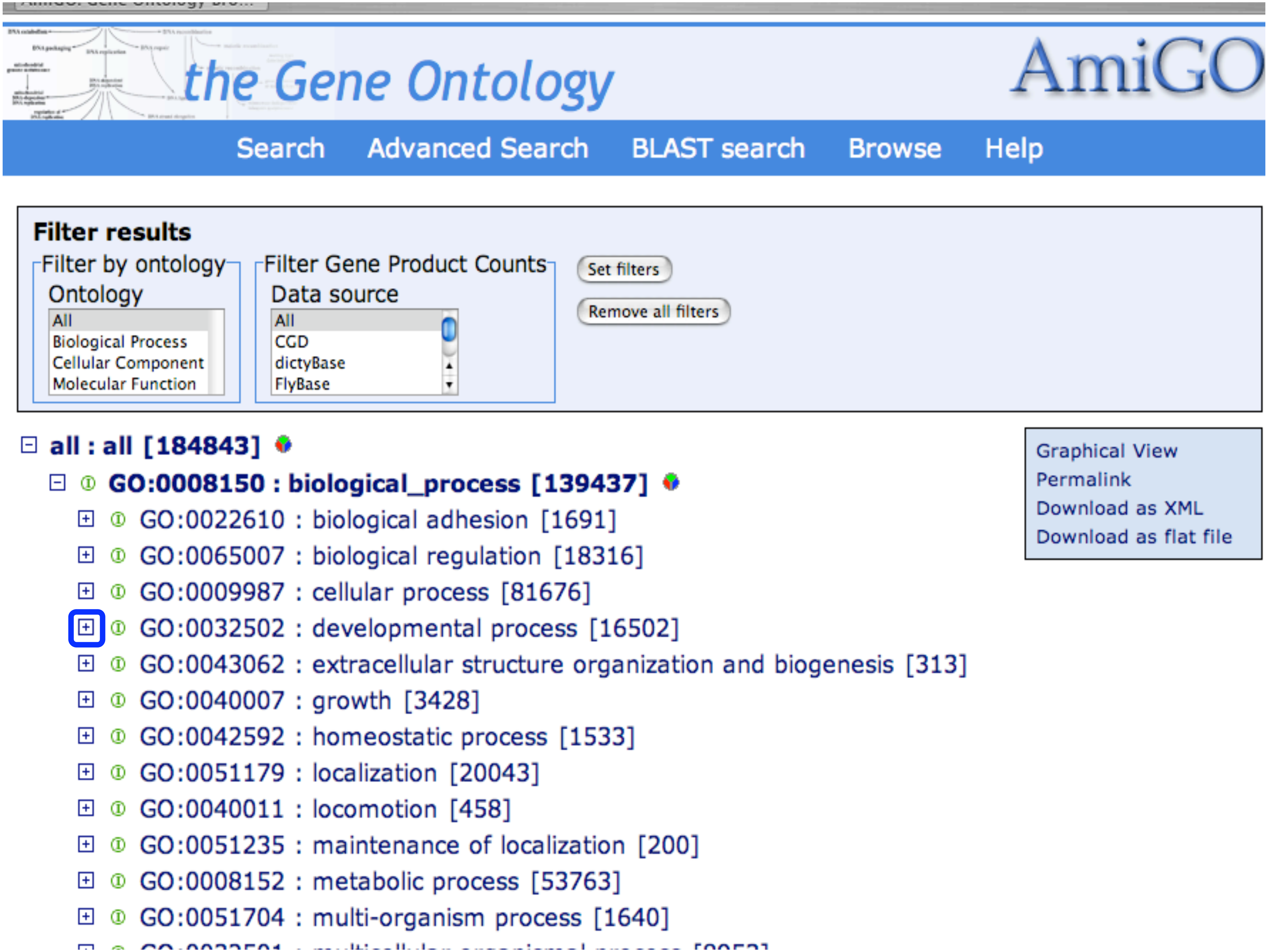
⊞ GO:0005575 : cellular_component [122434]

⊞ GO:0003674 : molecular_function [137219]

Last updated 2007-01-11

Terms of use • Disclaimer • GO helpdesk

Copyright © 1999-2006 the Gene Ontology Consortium



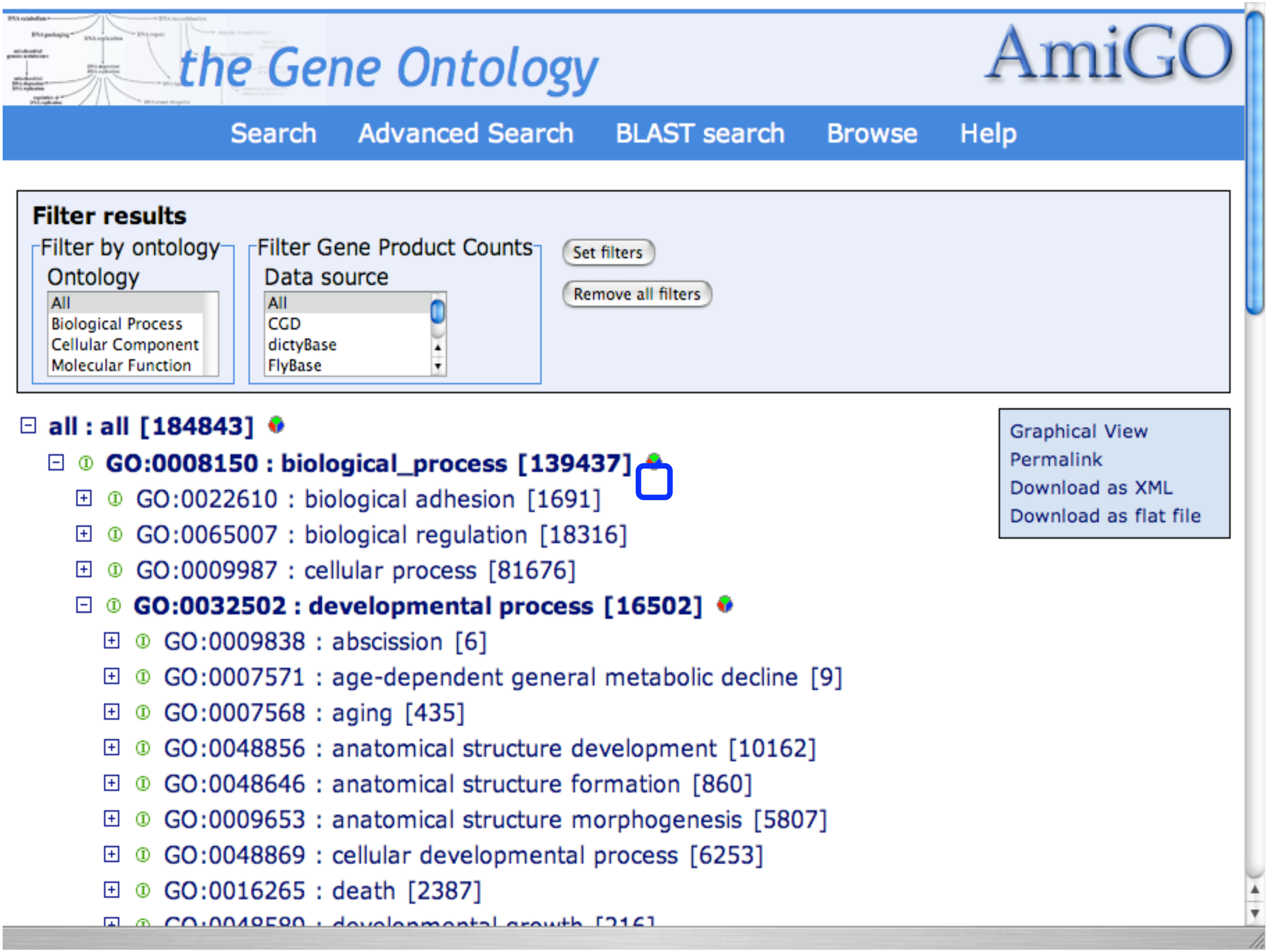
Filter results

Filter by ontology
Ontology
- All
- Biological Process
- Cellular Component
- Molecular Function

Filter Gene Product Counts
Data source
- All
- CGD
- dictyBase
- FlyBase

⊟ **all : all [184843]**

- ⊟ **GO:0008150 : biological_process [139437]**
  - ⊞ GO:0022610 : biological adhesion [1691]
  - ⊞ GO:0065007 : biological regulation [18316]
  - ⊞ GO:0009987 : cellular process [81676]
  - ⊞ GO:0032502 : developmental process [16502]
  - ⊞ GO:0043062 : extracellular structure organization and biogenesis [313]
  - ⊞ GO:0040007 : growth [3428]
  - ⊞ GO:0042592 : homeostatic process [1533]
  - ⊞ GO:0051179 : localization [20043]
  - ⊞ GO:0040011 : locomotion [458]
  - ⊞ GO:0051235 : maintenance of localization [200]
  - ⊞ GO:0008152 : metabolic process [53763]
  - ⊞ GO:0051704 : multi-organism process [1640]
  - ⊞ GO:0032501 : multicellular organismal process [8053]



Filter results

Filter by ontology

Ontology

- All
- Biological Process
- Cellular Component
- Molecular Function

Filter Gene Product Counts

Data source

- All
- CGD
- dictyBase
- FlyBase

⊟ **all : all [184843]**

⊟ **GO:0008150 : biological_process [139437]**

⊞ GO:0022610 : biological adhesion [1691]

⊞ GO:0065007 : biological regulation [18316]

⊞ GO:0009987 : cellular process [81676]

⊟ **GO:0032502 : developmental process [16502]**

⊞ GO:0009838 : abscission [6]

⊞ GO:0007571 : age-dependent general metabolic decline [9]

⊞ GO:0007568 : aging [435]

⊞ GO:0048856 : anatomical structure development [10162]

⊞ GO:0048646 : anatomical structure formation [860]

⊞ GO:0009653 : anatomical structure morphogenesis [5807]

⊞ GO:0048869 : cellular developmental process [6253]

⊞ GO:0016265 : death [2387]

⊞ GO:0048589 : developmental growth [216]

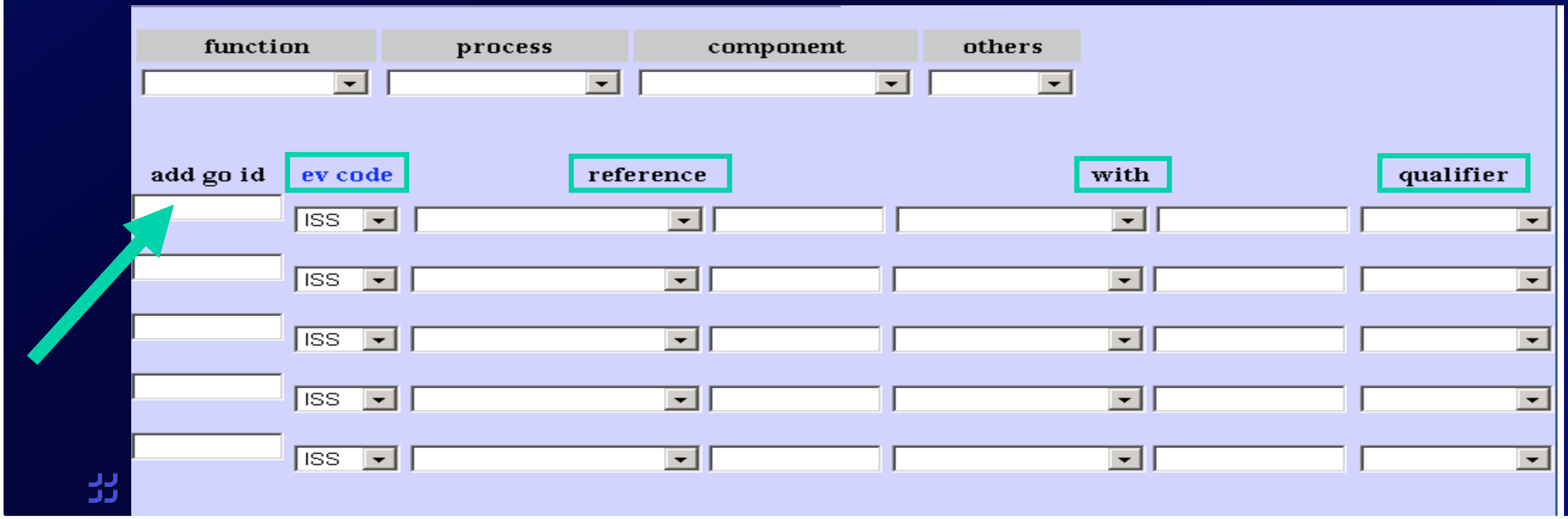
GO information to include

Independent of interface, add:

GO ID
Evidence code
Reference
Qualifier

The date is an important part of the annotation .

In Manatee:

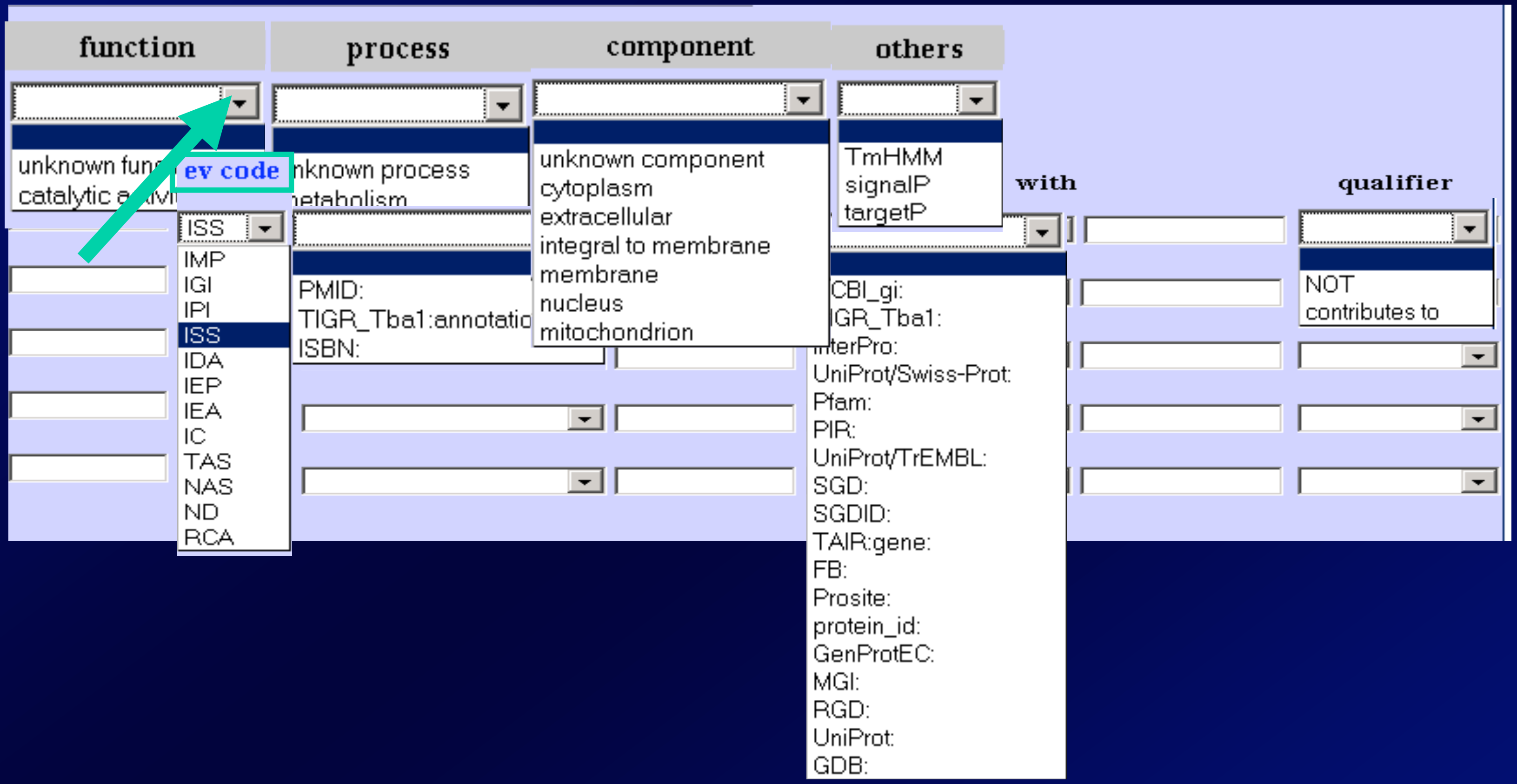


function | process | component | others

add go id | ev code | reference | with | qualifier

	ISS					
	ISS					
	ISS					
	ISS					
	ISS					

Filling in the GO information



Assigning a GO term

1) Read the literature, not just the abstract
2) Search for GO terms
3) Record the data

GO Annotations based on similarity

- Sequence or structure
  - Similarity to GO-annotated gene products
- Domains
- EC numbers
- Pathways
- Protein families

and many more…

http://www.geneontology.org/GO.indices.shtml

Annotating by similarity

use the evidence code ‘ISS’—inferred from sequence or structural similarity.

enter the database ID of the entity used to infer similarity in the ‘With’ field.

IEA: Inferred from Electronic Annotation

IEA is used when no curator has checked the annotation to verify its accuracy.

Use when an annotation:

- is based on "hits" in sequence similarity searches, if they have not been reviewed by curators
- is transferred from database records, if not reviewed by curators
- that depend directly on computation or automated transfer of annotations from a database.
  - The actual method used (BLAST search, SwissProt keyword mapping, etc.) doesn't matter.
  - If the method is match-based, a valid database ID *must* be entered in the with column.

GO Slim

- cut-down versions of the ontologies
- useful summary of GO annotation
- versions of GO Slims available
  - Eukaryotic GO slim
  - Plant GO slim
  - Yeast GO slim

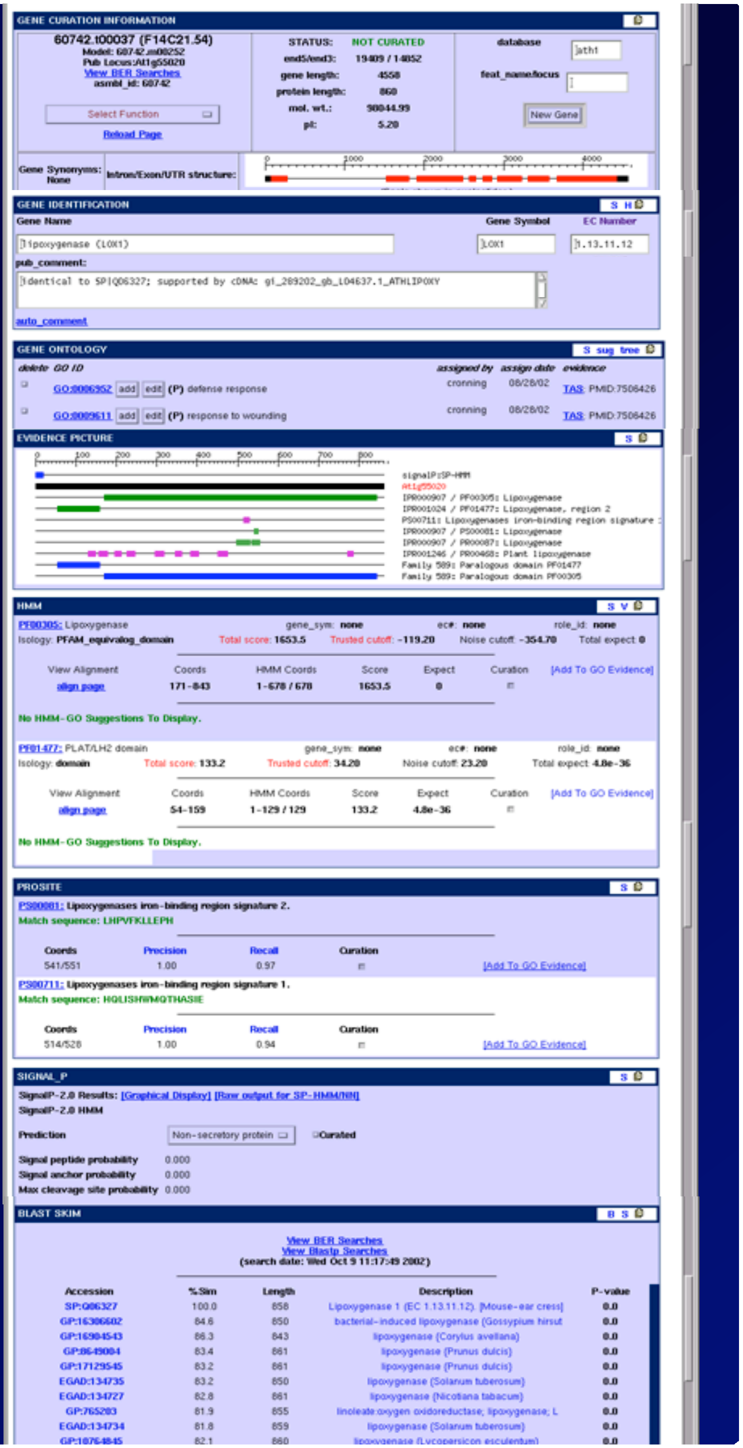
Points to remember

- GO enables querying across annotations
- The GO Consortium website has documentation and lists available tools
- AmiGO is available online and as downloadable resource
- GO Slims summarize your annotation
- GO annotations are worth the trouble—they enhance the value of research

MANATEE-

- Navigation, inspection & curation 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

STATUS:	NOT CURATED
end5/end3:	19409 / 14852
gene length:	4558
protein length:	860
mol. wt.:	98044.99
pI:	5.20

database: ath1
feat_name/locus:

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	(P) defense response	cronning	08/28/02	TAS: PMID:7508426
□	GO:0009611	(P) response to wounding	cronning	08/28/02	TAS: PMID:7508426

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	[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	858	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	86.3	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:10764845	82.1	860	lipoxygenase (Lycopersicon esculentum)	0.0