# Protein Families

João C. Setubal

University of São Paulo

Agosto 2012

# Motivation

- *Phytophthora Science* paper [Tyler et al., 2006]
- …Comparison of the [*P. sojae* and *P. ramorum*] genomes reveals a rapid expansion and diversification of many protein families associated with plant infection such as hydrolases, ABC transporters, protein toxins, proteinase inhibitors, and, in particular, a superfamily of 700 proteins with similarity to known oömycete avirulence genes.

# The concept of family

- A group where members share one or more characteristics
- In biology: usually descent or function
- Examples

# The family *felidae*

# Oomycota

- Kingdom: Chromalveolata
- Phylum: Heterokontophyta
- Class: **Oomycota**
- Orders (& families)
- Lagenidiales
  - Lagenidiaceae
  - Olpidiosidaceae
  - Sirolpidiaceae
- Leptomitales
  - Leptomitaceae
- Peronosporales
  - Albuginaceae
  - Peronosporaceae
  - Pythiaceae
- Rhipidiales
  - Rhipidaceae
- Saprolegniales
  - Ectrogellaceae
  - Haliphthoraceae
  - Leptolegniellaceae
  - Saprolegniaceae
- Thraustochytriales

Phytophthora

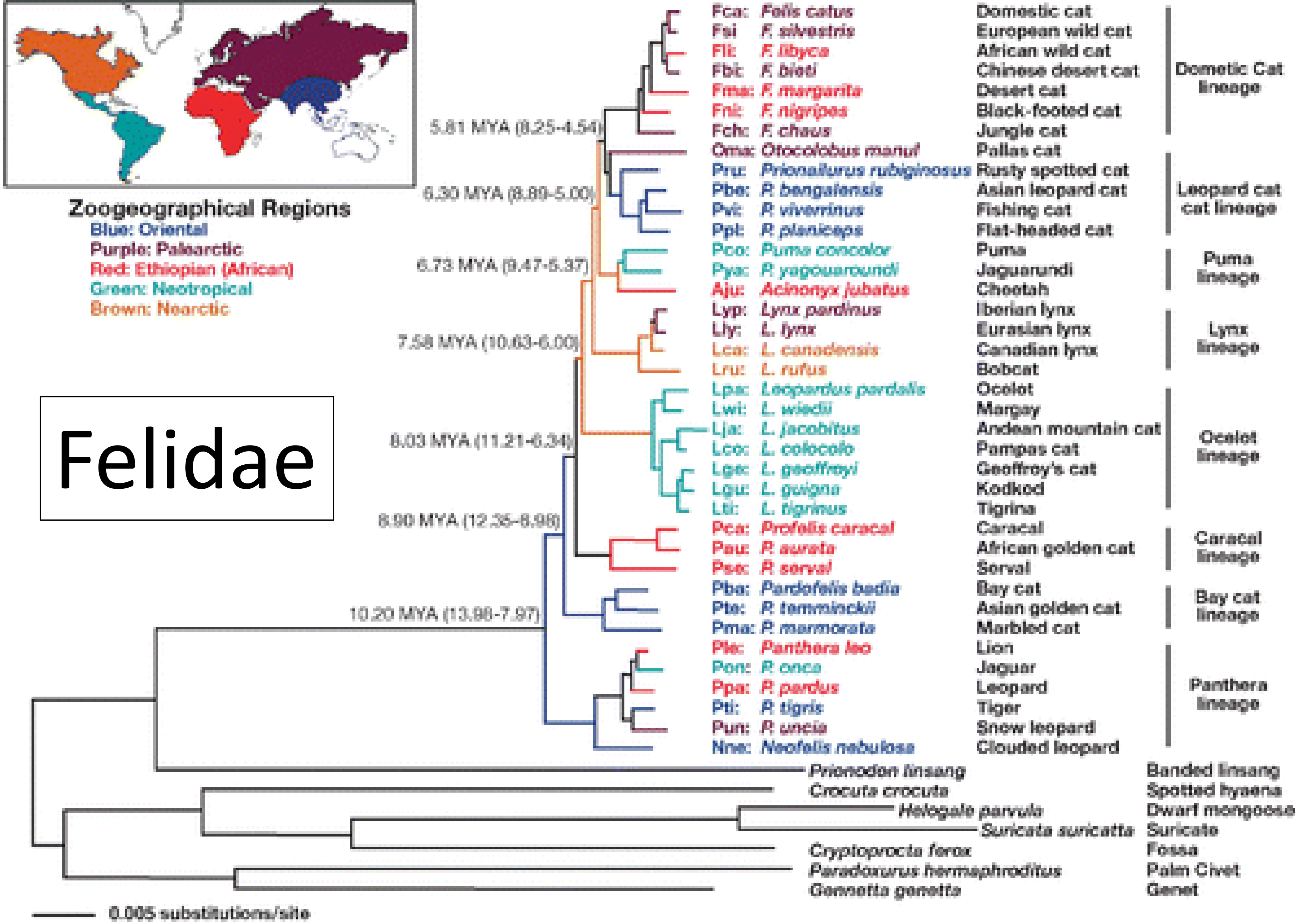
# Family by function

- The “family” of all effector proteins found in the phytophthora genus
- Members in general do not share an ancestor
- Function is ill-defined in this case

# Protein families

- Shared characteristic: common ancestor
- The concept of phylogeny is closely associated with that of family

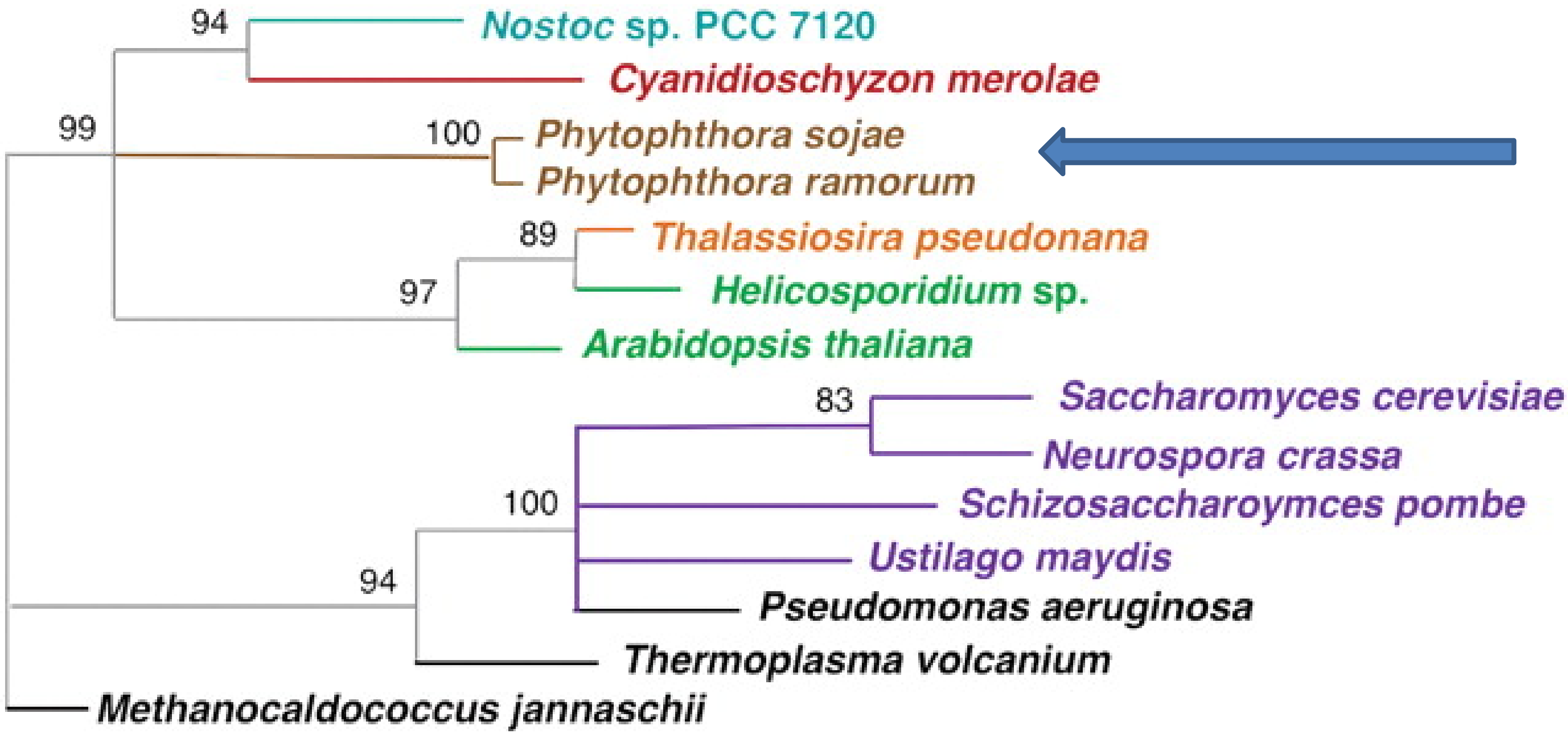
# Felidae



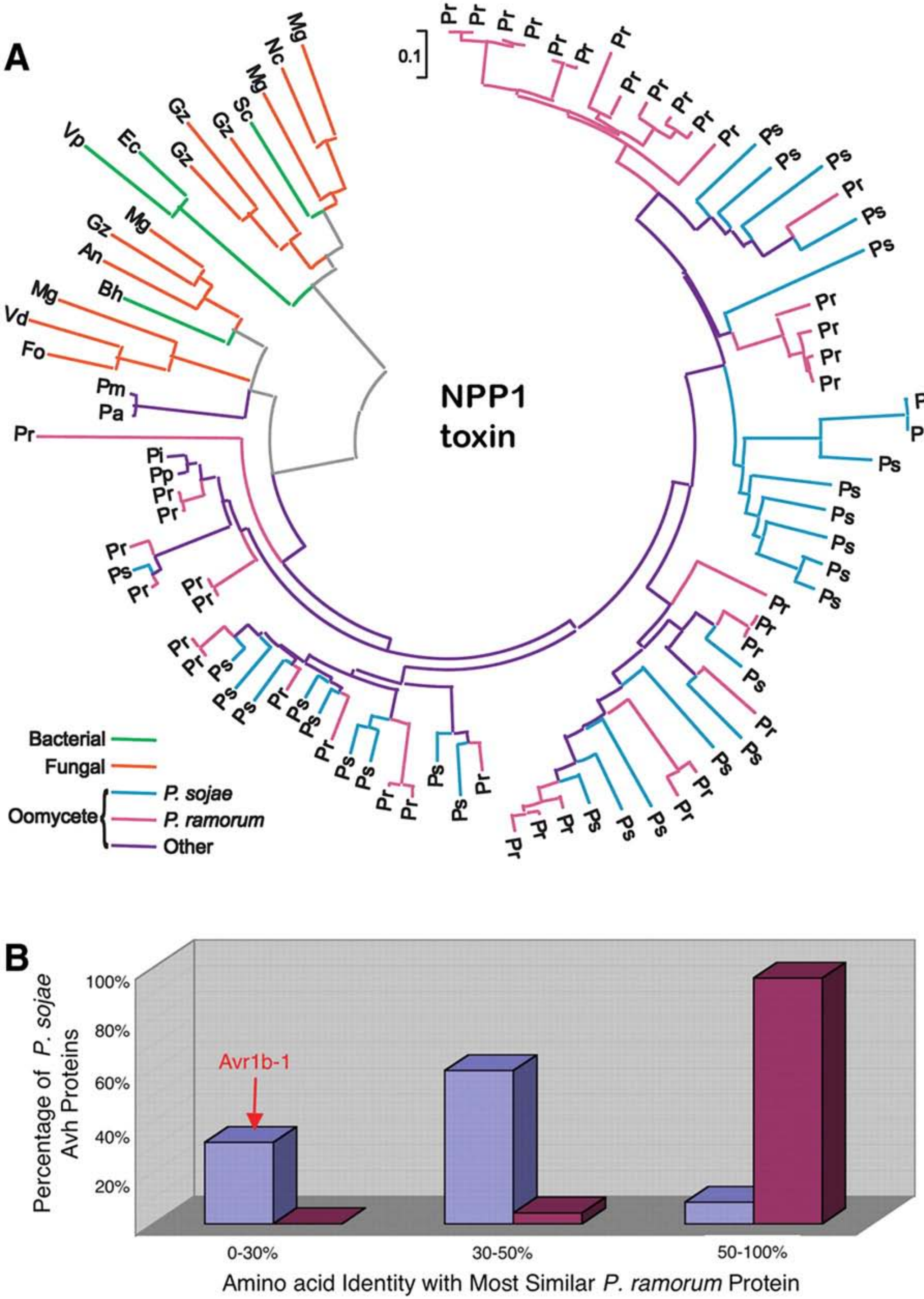
Zoogeographical Regions
Blue: Oriental
Purple: Palearctic
Red: Ethiopian (African)
Green: Neotropical
Brown: Nearctic
5.81 MYA (8.25-4.54)
6.30 MYA (8.89-5.00)
6.73 MYA (9.47-5.37)
7.58 MYA (10.63-6.00)
8.03 MYA (11.21-6.34)
8.90 MYA (12.35-6.98)
10.20 MYA (13.98-7.97)
Fca: Felis catus Domestic cat
Fsi F. silvestris European wild cat
Fli: F. libyca African wild cat
Fbi: F. bieti Chinese desert cat
Fma: F. margarita Desert cat
Fni: F. nigripes Black-footed cat
Fch: F. chaus Jungle cat
Oma: Otocolobus manul Pallas cat
Pru: Prionailurus rubiginosus Rusty spotted cat
Pbe: P. bengalensis Asian leopard cat
Pvi: P. viverrinus Fishing cat
Ppl: P. planiceps Flat-headed cat
Pco: Puma concolor Puma
Pya: P. yagouaroundi Jaguarundi
Aju: Acinonyx jubatus Cheetah
Lyp: Lynx pardinus Iberian lynx
Lly: L. lynx Eurasian lynx
Lca: L. canadensis Canadian lynx
Lru: L. rufus Bobcat
Lpa: Leopardus pardalis Ocelot
Lwi: L. wiedii Margay
Lja: L. jacobitus Andean mountain cat
Lco: L. colocolo Pampas cat
Lge: L. geoffroyi Geoffroy's cat
Lgu: L. guigna Kodkod
Lti: L. tigrinus Tigrina
Pca: Profelis caracal Caracal
Pau: P. aurata African golden cat
Pse: P. serval Serval
Pba: Pardofelis badia Bay cat
Pte: P. temminckii Asian golden cat
Pma: P. marmorata Marbled cat
Ple: Panthera leo Lion
Pon: P. onca Jaguar
Ppa: P. pardus Leopard
Pti: P. tigris Tiger
Pun: P. uncia Snow leopard
Nne: Neofelis nebulosa Clouded leopard
Dometic Cat lineage
Leopard cat cat lineage
Puma lineage
Lynx lineage
Ocelot lineage
Caracal lineage
Bay cat lineage
Panthera lineage
Prionodon linsang Banded linsang
Crocuta crocuta Spotted hyaena
Helogale parvula Dwarf mongoose
Suricata suricatta Suricate
Cryptoprocta ferox Fossa
Paradoxurus hermaphroditus Palm Civet
Genetta genetta Genet
0.005 substitutions/site

O'Brien SJ, Johnson WE. 2005.
Annu. Rev. Genomics Hum. Genet. 6:407–29

## 2-isopropylmalate synthase (leucine biosynthesis)



Tyler et al., *Science*, 2006



Tyler et al., *Science*, 2006

# Family (or grouping) by function

- A controlled vocabulary is necessary
- Gene Ontology

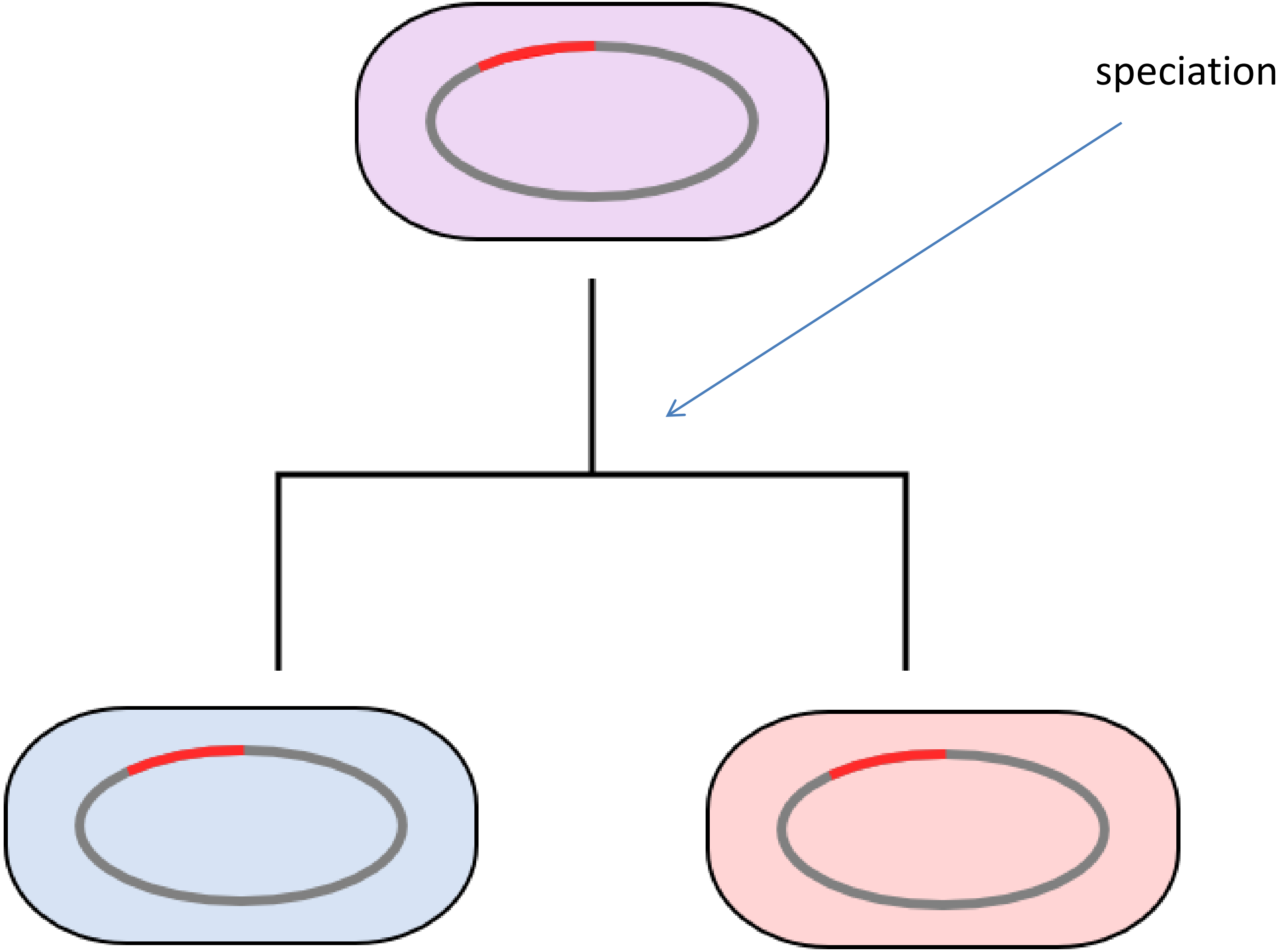
# Protein families by homology

- Aula 1
  - Important concepts
  - How to build a protein family
  - Protein family resources
- Aula 2
  - Phylogeny

# Homology

- Two genes that share an ancestor are said to be homologous
- Often (but wrongly) used with the sense of "similar"
- Similarity does not necessarily mean homology!
- Two kinds of homologous genes
  - Orthologs and paralogs

# Orthologs



# Paralogs

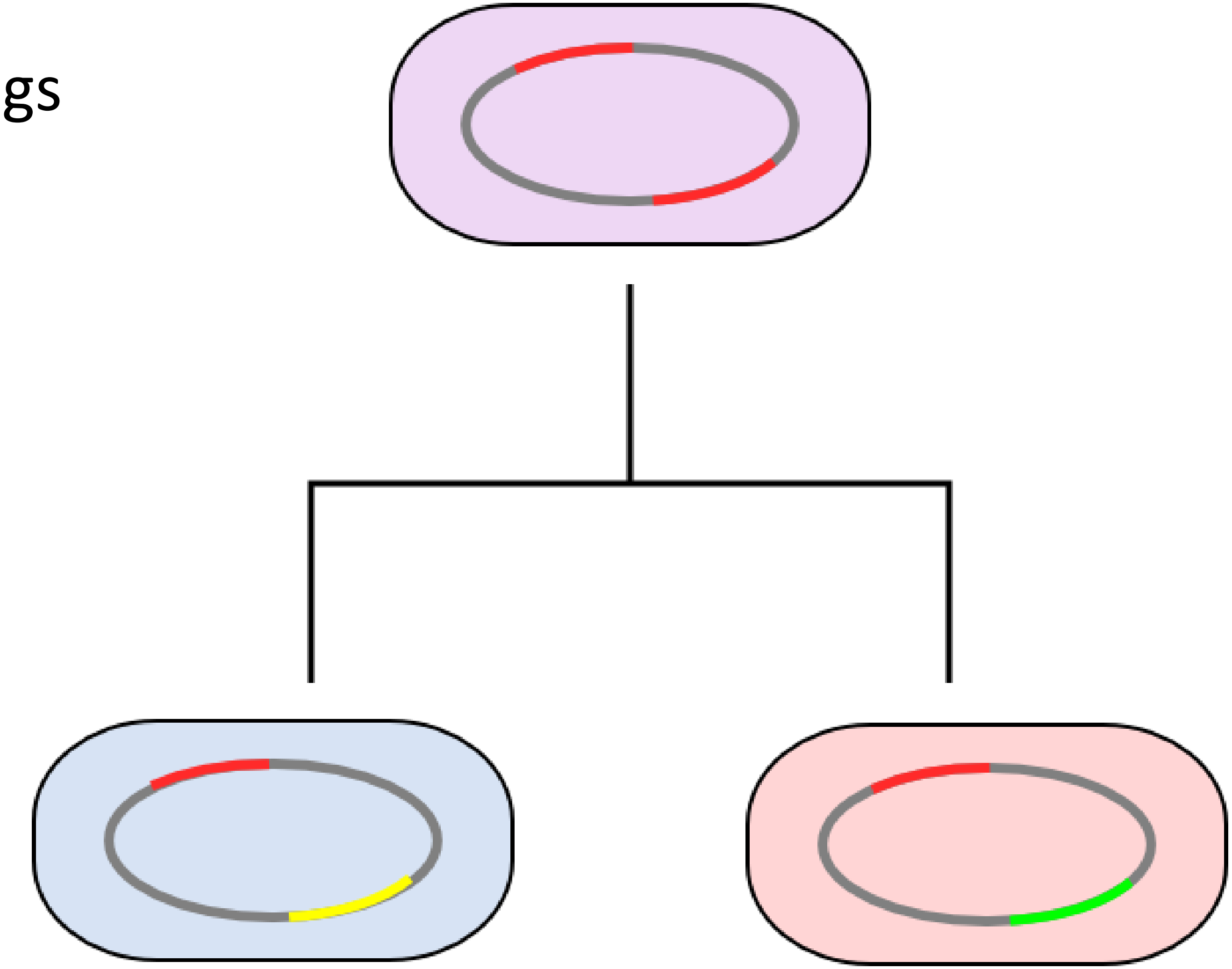


Figure by C. Lasher

# Protein family

- Operational definitions
  - Two proteins are in the same family if their genes are homologous
- or
  - Two proteins are in the same family if their genes are orthologous

- Will abuse language and mention homologous genes and homologous proteins

# In-paralogs

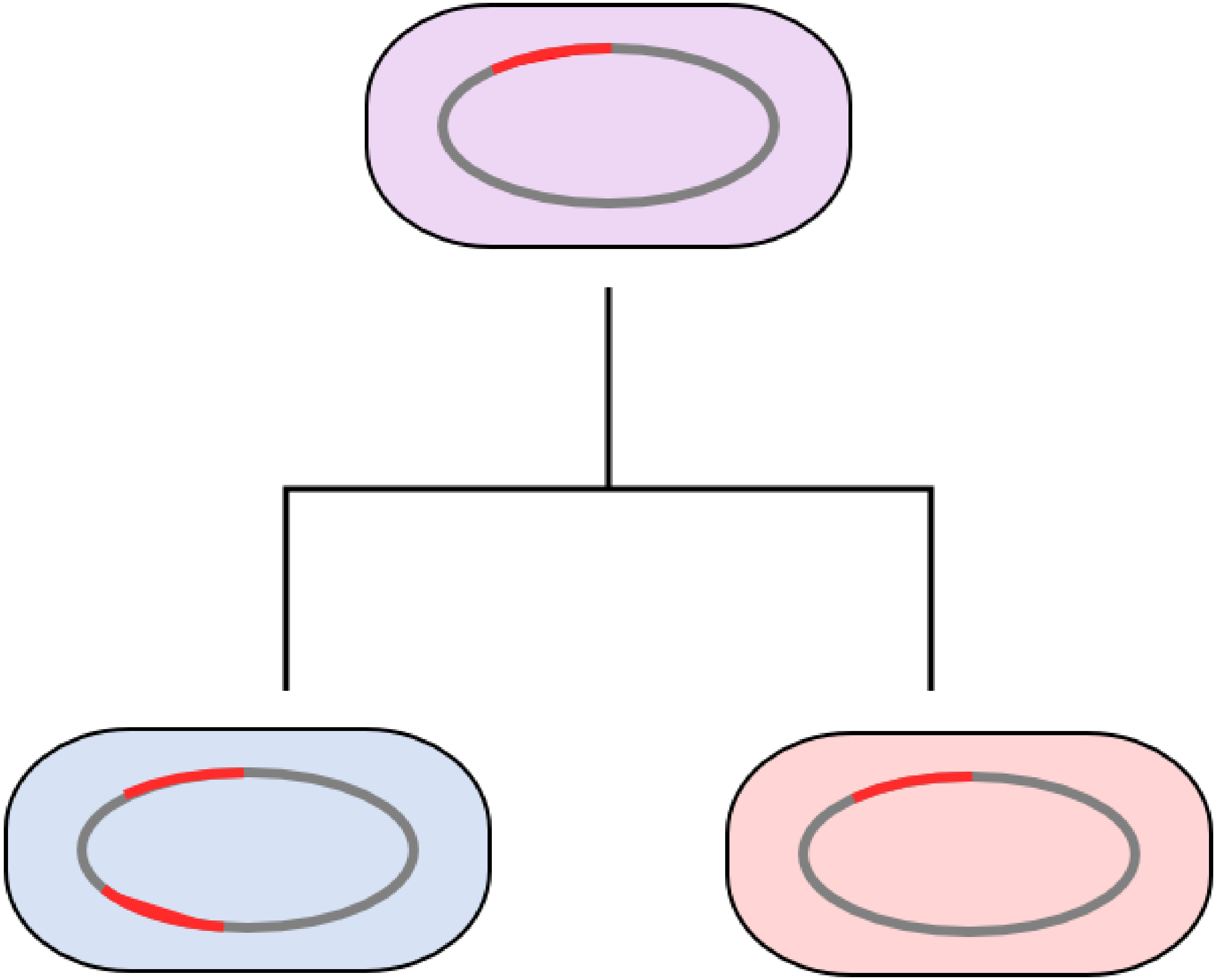
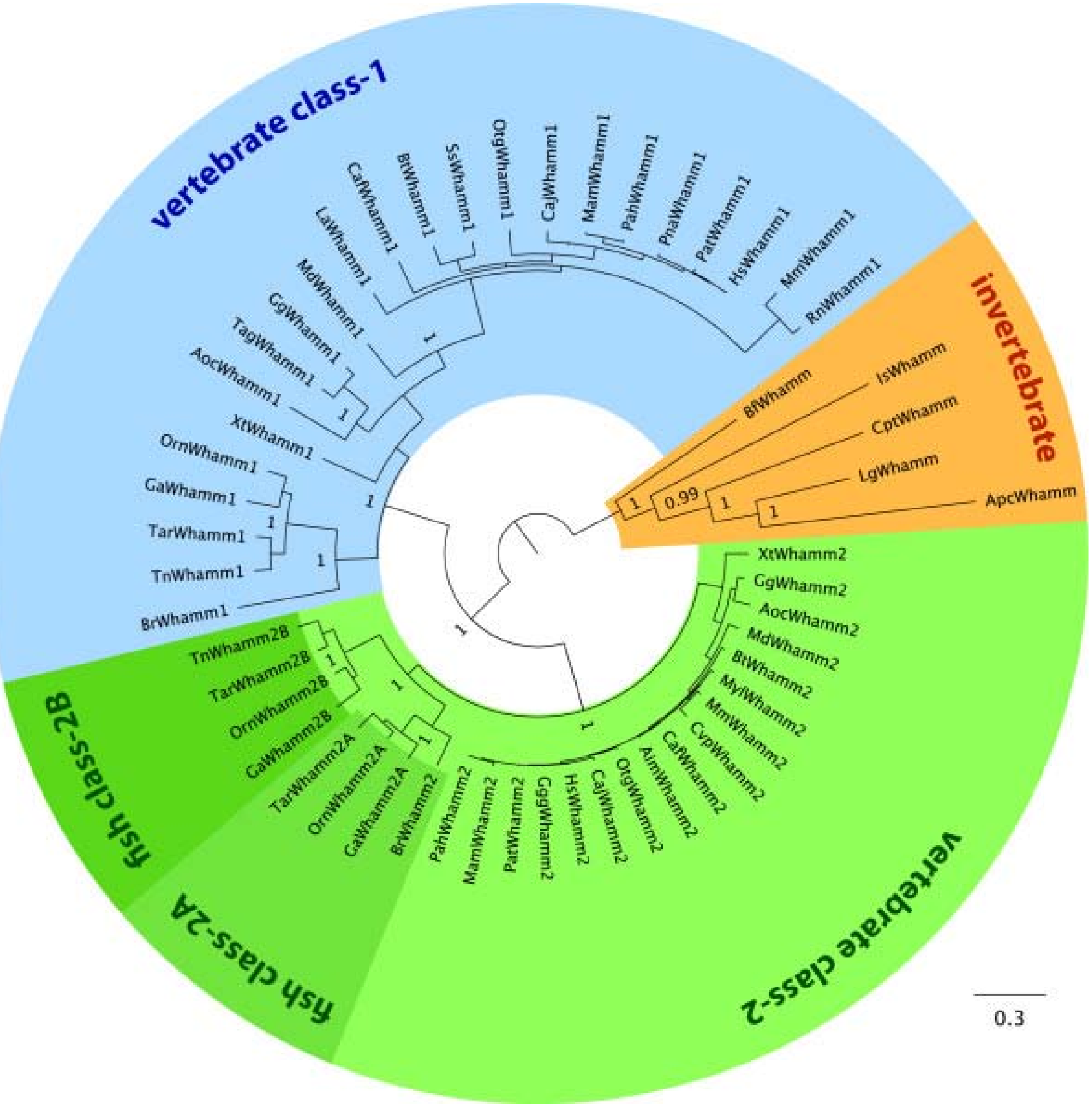


Figure by C. Lasher

# Homology and function

- We would like orthologous proteins to have the same function
- This is generally but not always the case
- Paralagous genes are more prone to develop new functions
  - Neo-functionalization
- In practice
  - Protein family: homology and shared function
  - Superfamilies and subfamilies



**Phylogenetic tree of the WHAMM proteins**

Kollmar *et al. BMC Research Notes* 2012 **5**:88 doi:10.1186/1756-0500-5-88

# Audiences for this lecture

1. You have a sequence and you want to build its family
2. You want to explore and use existing protein family resources for families you are interested in
3. You have a new genome and you want to place all of its genes into their families

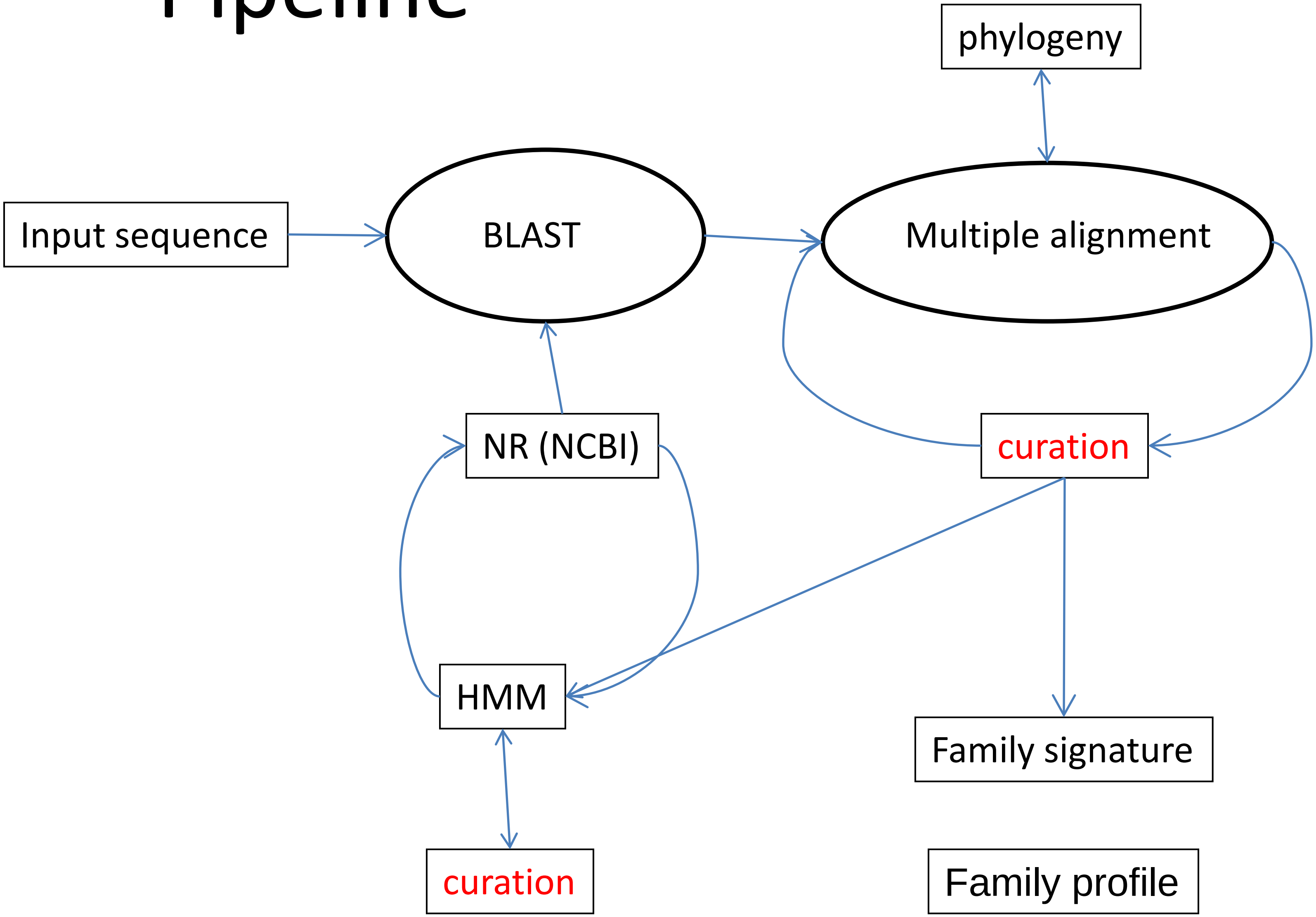
# Audience 1

You have a sequence and you want to build its family

# How to build a family

- Given a protein sequence
  - Determine other members
  - Create multiple alignment
  - Create family signature
  - Create model (Hidden Markov model)

# Pipeline



phylogeny

Input sequence

BLAST

Multiple alignment

NR (NCBI)

curation

HMM

Family signature

curation

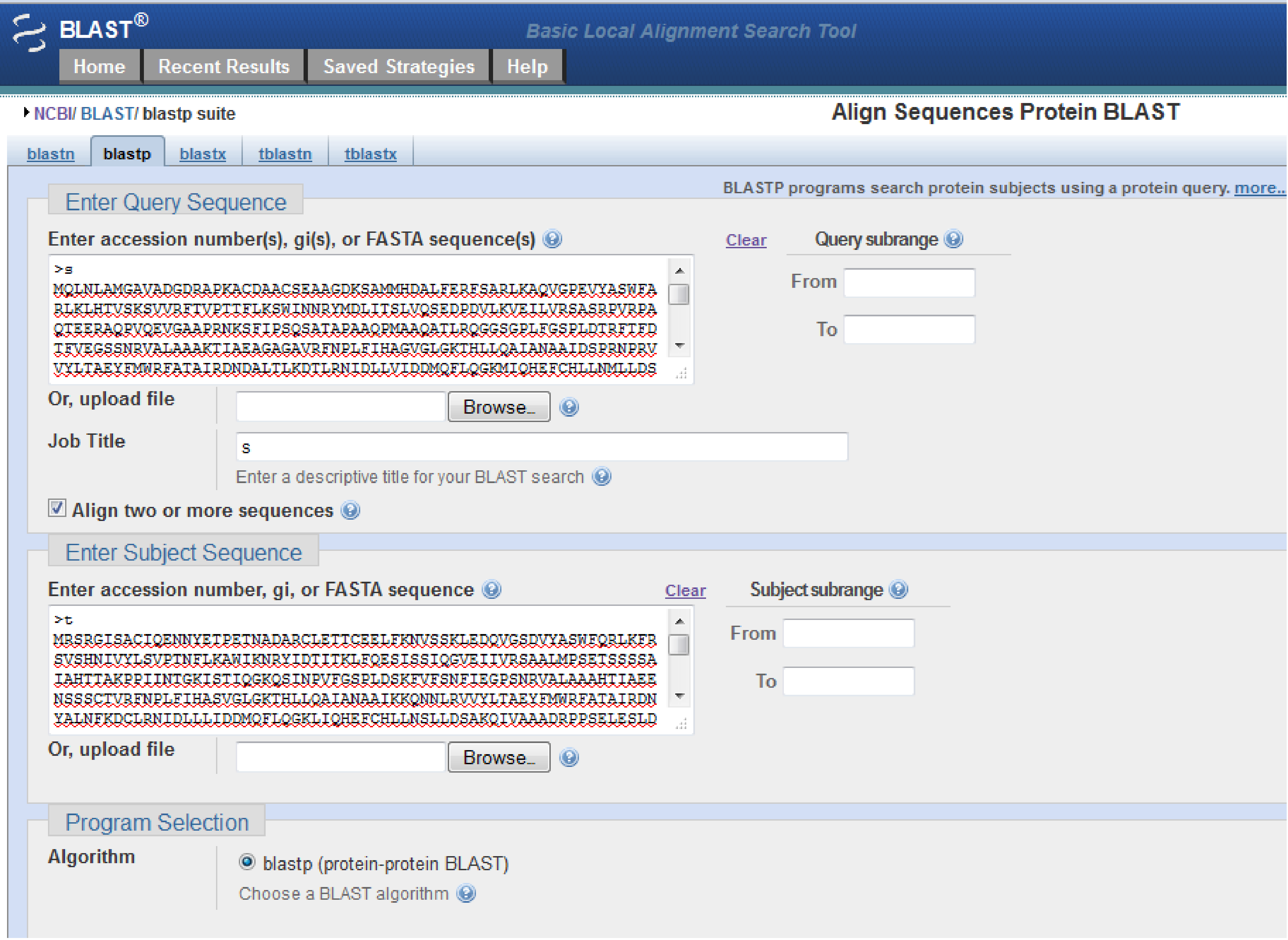
Family profile

# Problems & Tools

- BLAST
  - Max e-value and/or minimum identity
  - Minimum coverage
    - 80% query, 80% subject
- PSI-BLAST
- Multiple Alignment
  - Aula 2
- HMM
  - Use Pfam package
- If you don’t know what you’re doing, don’t try this at home! ☺

# Comparação de sequencias

- Similaridade “suficiente” → mesma função
- O que é similaridade?
- O que é “suficiente”?
- Google das sequencias: BLAST
- Basic Local Alignment Search Tool
- Altschul et al., 1990, 1997
- No ano 2000 já tinham mais de 13.000 citações



Align Sequences Protein BLAST

BLASTP programs search protein subjects using a protein query. more..

Enter Query Sequence

Enter accession number(s), gi(s), or FASTA sequence(s)

```
>s
MQLNLAMGAVADGDRAPKACDAACSEAAGDKSAMMHDALFERFSARLKAQVGPEVYASWFA
RLKLHTVSKSVVRFTVPTTFLKSWINNRYMDLITSLVQSEDPDVLKVEILVRSASRPVRPA
QTEERAQPVQEVGAAPRNKSFIPSQSATAPAAQPMAAQATLRQGGSGPLFGSPLDTRFTFD
TFVEGSSNRVALAAAKTIAEAGAGAVRFNPLFIHAGVGLGKTHLLQAIANAAIDSPRNPRV
VYLTAEYFMWRFATAIRDNDALTLKDTLRNIDLLVIDDMQFLQGKMIQHEFCHLLNMLLDS
```

Query subrange

From

To

Or, upload file

Job Title: s

Enter a descriptive title for your BLAST search

☑ Align two or more sequences

Enter Subject Sequence

Enter accession number, gi, or FASTA sequence

```
>t
MRSRGISACIQENNYETPETNADARCLETTCEELFKNVSSKLEDQVGSDVYASWFQRLKFR
SVSHNIVYLSVPTNFLKAWIKNRYIDTITKLFQESISSIQGVEIIVRSAALMPSETSSSSA
IAHTTAKPPIINTGKISTIQGKQSINPVFGSPLDSKFVFSNFIEGPSNRVALAAAHTIAEE
NSSSCTVRFNPLFIHASVGLGKTHLLQAIANAAIKKQNNLRVVYLTAEYFMWRFATAIRDN
YALNFKDCLRNIDLLLIDDMQFLQGKLIQHEFCHLLNSLLDSAKQIVAAADRPPSELESLD
```

Subject subrange

From

To

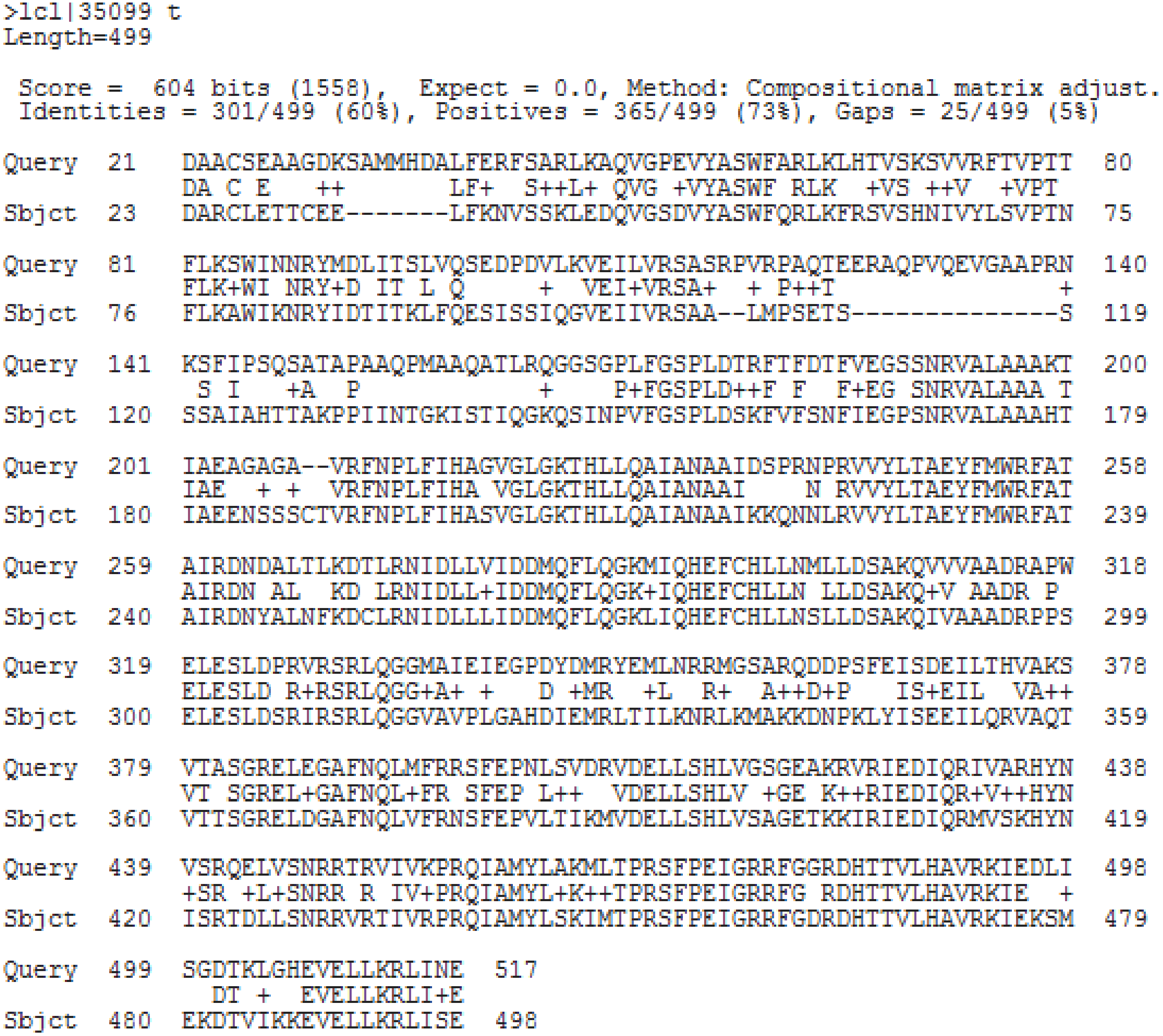
Or, upload file

Program Selection

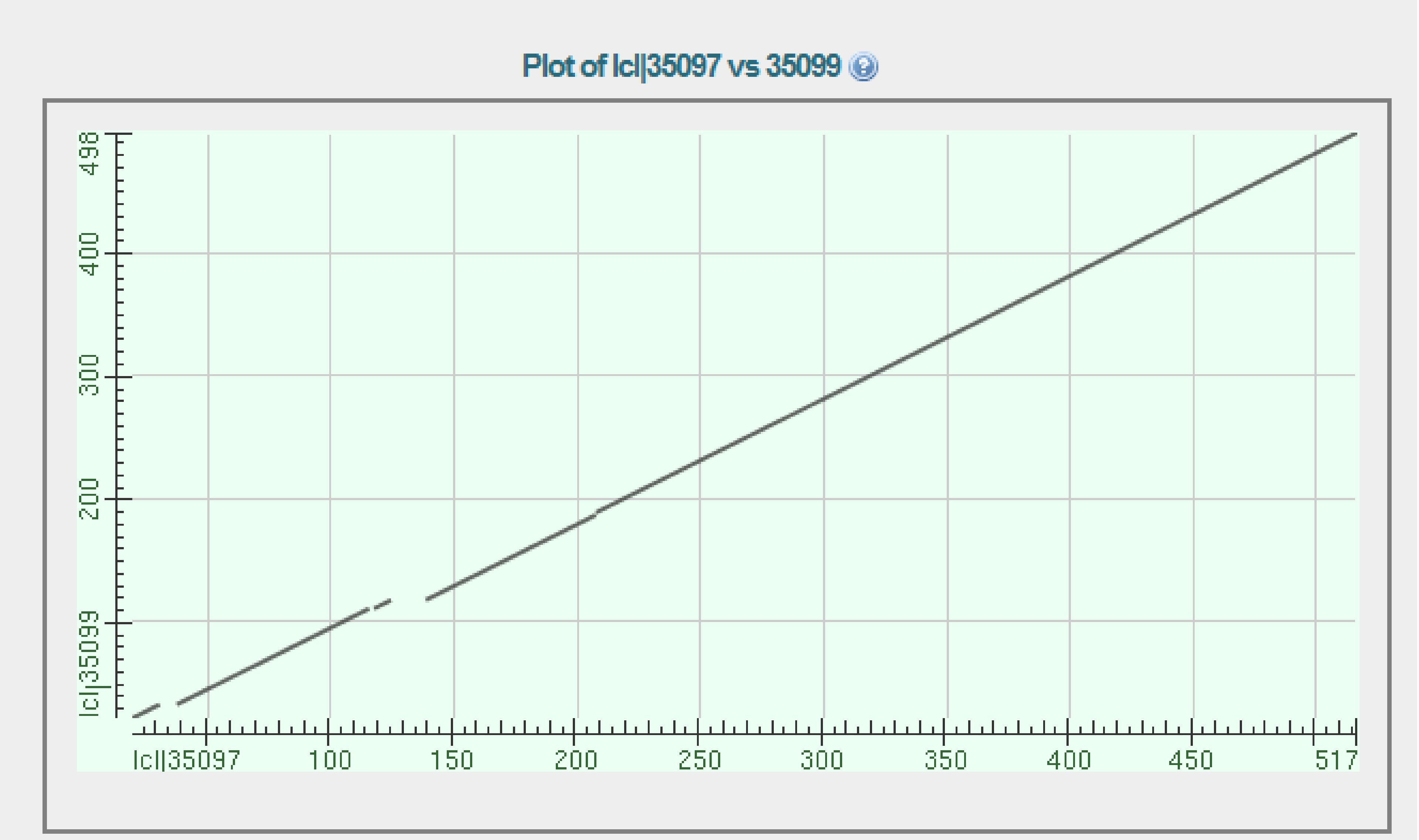
Algorithm

● blastp (protein-protein BLAST)

Choose a BLAST algorithm

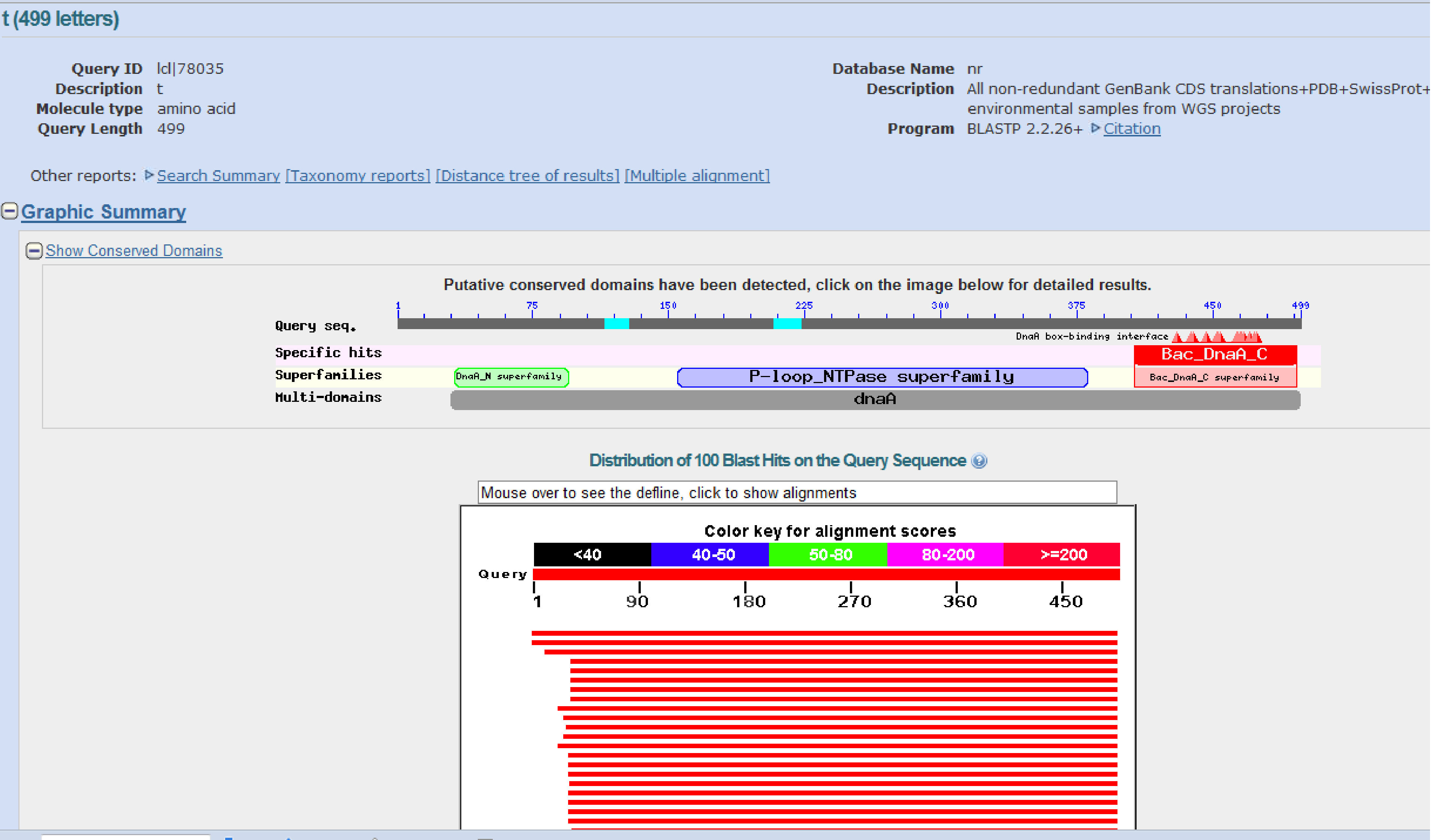


```
>lcl|35099 t
Length=499

 Score =  604 bits (1558),  Expect = 0.0, Method: Compositional matrix adjust.
 Identities = 301/499 (60%), Positives = 365/499 (73%), Gaps = 25/499 (5%)

Query  21   DAACSEAAGDKSAMMHDALFERFSARLKAQVGPEVYASWFARLKLHTVSKSVVRFTVPTT  80
            DA C E   ++       LF+  S++L+ QVG +VYASWF RLK  +VS ++V  +VPT
Sbjct  23   DARCLETTCEE-------LFKNVSSKLEDQVGSDVYASWFQRLKFRSVSHNIVYLSVPTN  75

Query  81   FLKSWINNRYMDLITSLVQSEDPDVLKVEILVRSASRPVRPAQTEERAQPVQEVGAAPRN  140
            FLK+WI NRY+D IT L Q     +  VEI+VRSA+  + P++T               +
Sbjct  76   FLKAWIKNRYIDTITKLFQESISSIQGVEIIVRSAA--LMPSETS--------------S  119

Query  141  KSFIPSQSATAPAAQPMAAQATLRQGGSGPLFGSPLDTRFTFDTFVEGSSNRVALAAAKT  200
             S I   +A  P           +    P+FGSPLD++F F  F+EG SNRVALAAA T
Sbjct  120  SSAIAHTTAKPPIINTGKISTIQGKQSINPVFGSPLDSKFVFSNFIEGPSNRVALAAAHT  179

Query  201  IAEAGAGA--VRFNPLFIHAGVGLGKTHLLQAIANAAIDSPRNPRVVYLTAEYFMWRFAT  258
            IAE  + +  VRFNPLFIHA VGLGKTHLLQAIANAAI    N RVVYLTAEYFMWRFAT
Sbjct  180  IAEENSSSCTVRFNPLFIHASVGLGKTHLLQAIANAAIKKQNNLRVVYLTAEYFMWRFAT  239

Query  259  AIRDNDALTLKDTLRNIDLLVIDDMQFLQGKMIQHEFCHLLNMLLDSAKQVVVAADRAPW  318
            AIRDN AL  KD LRNIDLL+IDDMQFLQGK+IQHEFCHLLN LLDSAKQ+V AADR P
Sbjct  240  AIRDNYALNFKDCLRNIDLLLIDDMQFLQGKLIQHEFCHLLNSLLDSAKQIVAAADRPPS  299

Query  319  ELESLDPRVRSRLQGGMAIEIEGPDYDMRYEMLNRRMGSARQDDPSFEISDEILTHVAKS  378
            ELESLD R+RSRLQGG+A+ +   D +MR  +L  R+  A++D+P   IS+EIL  VA++
Sbjct  300  ELESLDSRIRSRLQGGVAVPLGAHDIEMRLTILKNRLKMAKKDNPKLYISEEILQRVAQT  359

Query  379  VTASGRELEGAFNQLMFRRSFEPNLSVDRVDELLSHLVGSGEAKRVRIEDIQRIVARHYN  438
            VT SGREL+GAFNQL+FR SFEP L++  VDELLSHLV +GE K++RIEDIQR+V++HYN
Sbjct  360  VTTSGRELDGAFNQLVFRNSFEPVLTIKMVDELLSHLVSAGETKKIRIEDIQRMVSKHYN  419

Query  439  VSRQELVSNRRTRVIVKPRQIAMYLAKMLTPRSFPEIGRRFGGRDHTTVLHAVRKIEDLI  498
            +SR +L+SNRR R IV+PRQIAMYL+K++TPRSFPEIGRRFG RDHTTVLHAVRKIE  +
Sbjct  420  ISRTDLLSNRRVRTIVRPRQIAMYLSKIMTPRSFPEIGRRFGDRDHTTVLHAVRKIEKSM  479

Query  499  SGDTKLGHEVELLKRLINE  517
              DT +  EVELLKRLI+E
Sbjct  480  EKDTVIKKEVELLKRLISE  498
```



Plot of lcl|35097 vs 35099

# Buscando no GenBank



t (499 letters)

Query ID lcl|78035
Description t
Molecule type amino acid
Query Length 499

Database Name nr
Description All non-redundant GenBank CDS translations+PDB+SwissProt+
environmental samples from WGS projects
Program BLASTP 2.2.26+ ▷ Citation

Other reports: ▷ Search Summary [Taxonomy reports] [Distance tree of results] [Multiple alignment]

⊟ Graphic Summary

⊟ Show Conserved Domains

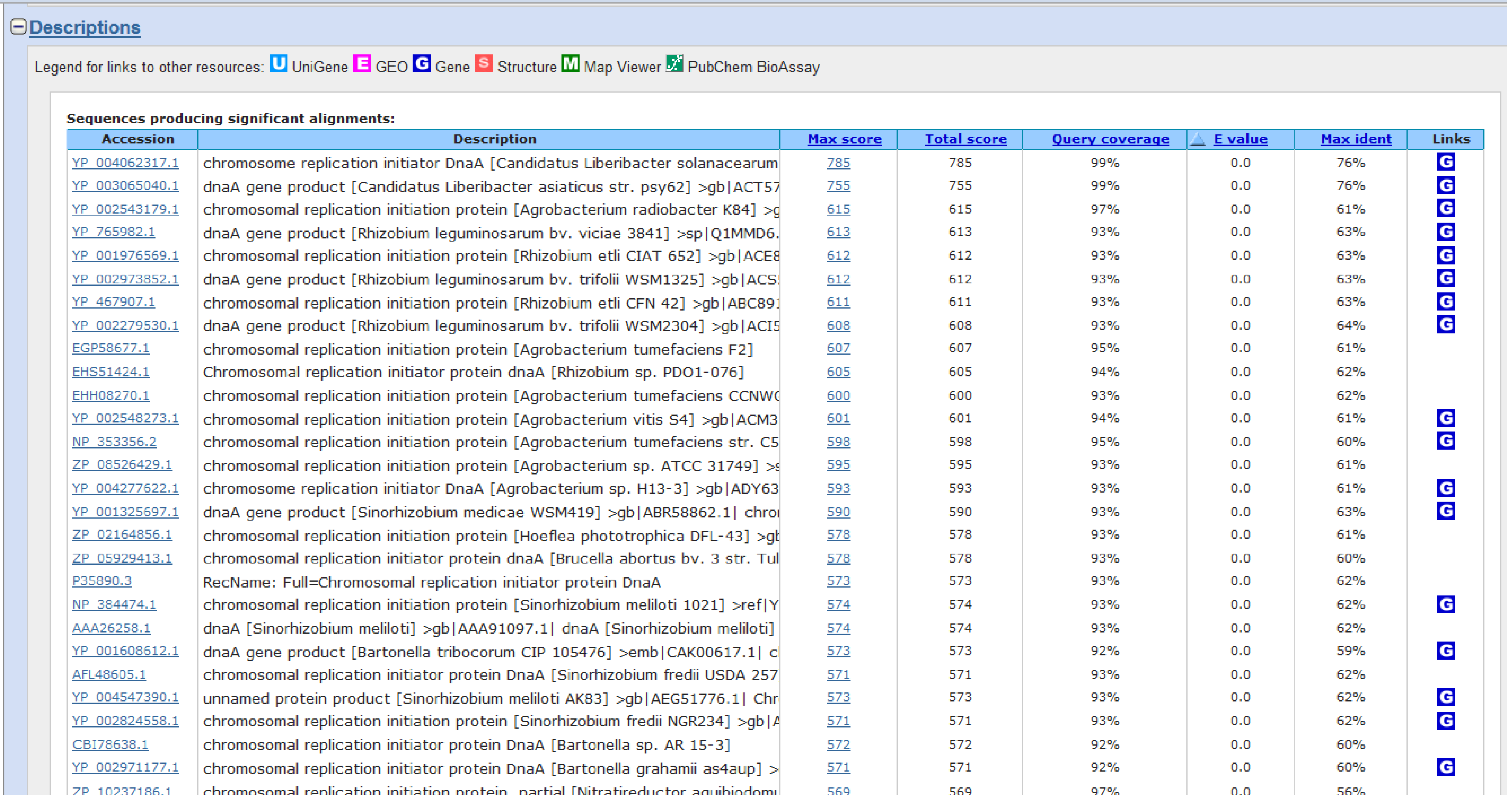
Putative conserved domains have been detected, click on the image below for detailed results.

Distribution of 100 Blast Hits on the Query Sequence

Mouse over to see the define, click to show alignments

Color key for alignment scores

# Lista de hits



⊖ Descriptions

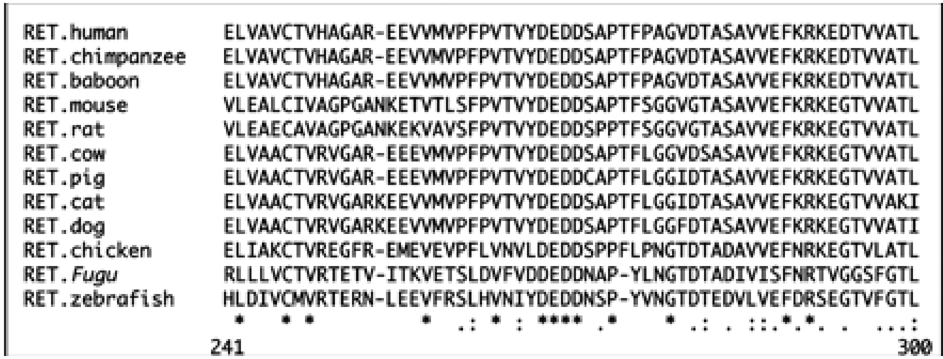
Legend for links to other resources: U UniGene E GEO G Gene S Structure M Map Viewer PubChem BioAssay

**Sequences producing significant alignments:**

| Accession | Description | Max score | Total score | Query coverage | △ E value | Max ident | Links |
|---|---|---|---|---|---|---|---|
| YP_004062317.1 | chromosome replication initiator DnaA [Candidatus Liberibacter solanacearum | 785 | 785 | 99% | 0.0 | 76% | G |
| YP_003065040.1 | dnaA gene product [Candidatus Liberibacter asiaticus str. psy62] >gb\|ACT57 | 755 | 755 | 99% | 0.0 | 76% | G |
| YP_002543179.1 | chromosomal replication initiation protein [Agrobacterium radiobacter K84] >g | 615 | 615 | 97% | 0.0 | 61% | G |
| YP_765982.1 | dnaA gene product [Rhizobium leguminosarum bv. viciae 3841] >sp\|Q1MMD6. | 613 | 613 | 93% | 0.0 | 63% | G |
| YP_001976569.1 | chromosomal replication initiation protein [Rhizobium etli CIAT 652] >gb\|ACE8 | 612 | 612 | 93% | 0.0 | 63% | G |
| YP_002973852.1 | dnaA gene product [Rhizobium leguminosarum bv. trifolii WSM1325] >gb\|ACS | 612 | 612 | 93% | 0.0 | 63% | G |
| YP_467907.1 | chromosomal replication initiation protein [Rhizobium etli CFN 42] >gb\|ABC89 | 611 | 611 | 93% | 0.0 | 63% | G |
| YP_002279530.1 | dnaA gene product [Rhizobium leguminosarum bv. trifolii WSM2304] >gb\|ACI5 | 608 | 608 | 93% | 0.0 | 64% | G |
| EGP58677.1 | chromosomal replication initiation protein [Agrobacterium tumefaciens F2] | 607 | 607 | 95% | 0.0 | 61% | |
| EHS51424.1 | Chromosomal replication initiator protein dnaA [Rhizobium sp. PDO1-076] | 605 | 605 | 94% | 0.0 | 62% | |
| EHH08270.1 | chromosomal replication initiation protein [Agrobacterium tumefaciens CCNWC | 600 | 600 | 93% | 0.0 | 62% | |
| YP_002548273.1 | chromosomal replication initiation protein [Agrobacterium vitis S4] >gb\|ACM3 | 601 | 601 | 94% | 0.0 | 61% | G |
| NP_353356.2 | chromosomal replication initiation protein [Agrobacterium tumefaciens str. C5 | 598 | 598 | 95% | 0.0 | 60% | G |
| ZP_08526429.1 | chromosomal replication initiation protein [Agrobacterium sp. ATCC 31749] >s | 595 | 595 | 93% | 0.0 | 61% | |
| YP_004277622.1 | chromosome replication initiator DnaA [Agrobacterium sp. H13-3] >gb\|ADY63 | 593 | 593 | 93% | 0.0 | 61% | G |
| YP_001325697.1 | dnaA gene product [Sinorhizobium medicae WSM419] >gb\|ABR58862.1\| chro | 590 | 590 | 93% | 0.0 | 63% | G |
| ZP_02164856.1 | chromosomal replication initiation protein [Hoeflea phototrophica DFL-43] >gb | 578 | 578 | 93% | 0.0 | 61% | |
| ZP_05929413.1 | chromosomal replication initiator protein dnaA [Brucella abortus bv. 3 str. Tul | 578 | 578 | 93% | 0.0 | 60% | |
| P35890.3 | RecName: Full=Chromosomal replication initiator protein DnaA | 573 | 573 | 93% | 0.0 | 62% | |
| NP_384474.1 | chromosomal replication initiation protein [Sinorhizobium meliloti 1021] >ref\|Y | 574 | 574 | 93% | 0.0 | 62% | G |
| AAA26258.1 | dnaA [Sinorhizobium meliloti] >gb\|AAA91097.1\| dnaA [Sinorhizobium meliloti] | 574 | 574 | 93% | 0.0 | 62% | |
| YP_001608612.1 | dnaA gene product [Bartonella tribocorum CIP 105476] >emb\|CAK00617.1\| c | 573 | 573 | 92% | 0.0 | 59% | G |
| AFL48605.1 | chromosomal replication initiator protein DnaA [Sinorhizobium fredii USDA 257 | 571 | 571 | 93% | 0.0 | 62% | |
| YP_004547390.1 | unnamed protein product [Sinorhizobium meliloti AK83] >gb\|AEG51776.1\| Chr | 573 | 573 | 93% | 0.0 | 62% | G |
| YP_002824558.1 | chromosomal replication initiation protein [Sinorhizobium fredii NGR234] >gb\|A | 571 | 571 | 93% | 0.0 | 62% | G |
| CBI78638.1 | chromosomal replication initiator protein DnaA [Bartonella sp. AR 15-3] | 572 | 572 | 92% | 0.0 | 60% | |
| YP_002971177.1 | chromosomal replication initiator protein DnaA [Bartonella grahamii as4aup] > | 571 | 571 | 92% | 0.0 | 60% | G |
| ZP_10237186.1 | chromosomal replication initiation protein, partial [Nitratireductor aquibiodomu | 569 | 569 | 97% | 0.0 | 56% | |

# Outros tipos de análise

- Alinhamento múltiplo



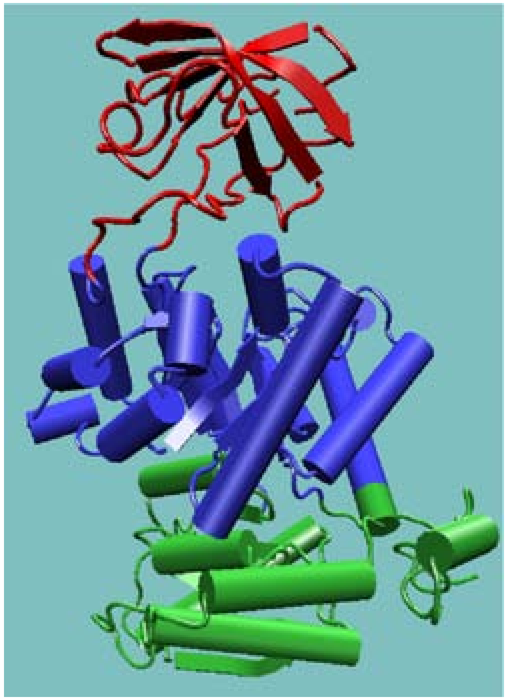
```
RET.human         ELVAVCTVHAGAR-EEVVMVPFPVTVYDEDDSAPTFPAGVDTASAVVEFKRKEDTVVATL
RET.chimpanzee    ELVAVCTVHAGAR-EEVVMVPFPVTVYDEDDSAPTFPAGVDTASAVVEFKRKEDTVVATL
RET.baboon        ELVAVCTVHAGAR-EEVVMVPFPVTVYDEDDSAPTFPAGVDTASAVVEFKRKEDTVVATL
RET.mouse         VLEALCIVAGPGANKETVTLSFPVTVYDEDDSAPTFSGGVGTASAVVEFKRKEGTVVATL
RET.rat           VLEAECAVAGPGANKEKVAVSFPVTVYDEDDSPPTFSGGVGTASAVVEFKRKEGTVVATL
RET.cow           ELVAACTVRVGAR-EEEVMVPFPVTVYDEDDSAPTFLGGVDSASAVVEFKRKEGTVVATL
RET.pig           ELVAACTVRVGAR-EEEVMVPFPVTVYDEDDCAPTFLGGIDTASAVVEFKRKEGTVVATL
RET.cat           ELVAACTVRVGARKEEVVMVPFPVTVYDEDDSAPTFLGGIDTASAVVEFKRKEGTVVAKI
RET.dog           ELVAACTVRVGARKEEVVMVPFPVTVYDEDDSAPTFLGGFDTASAVVEFKRKEGTVVATI
RET.chicken       ELIAKCTVREGFR-EMEVEVPFLVNVLDEDDSPPFLPNGTDTADAVVEFNRKEGTVLATL
RET.Fugu          RLLLVCTVRTETV-ITKVETSLDVFVDDEDDNAP-YLNGTDTADIVISFNRTVGGSFGTL
RET.zebrafish     HLDIVCMVRTERN-LEEVFRSLHVNIYDEDDNSP-YVNGTDTEDVLVEFDRSEGTVFGTL
                   *   * *         *  .: * : **** .*    * .: . ::.*.*. .  ...:
                 241                                                        300
```

- Construção de árvore filogenética
- Construção de uma assinatura de uma família
  - Modelo oculto de Markov (HMM)
- Predição de estrutura

*PNAS June 21, 2005 vol. 102 no. 25 8949-8954*

# Issues

- Two proteins may share a domain and be unrelated
  - BLAST false positives
- Multiple domain architectures

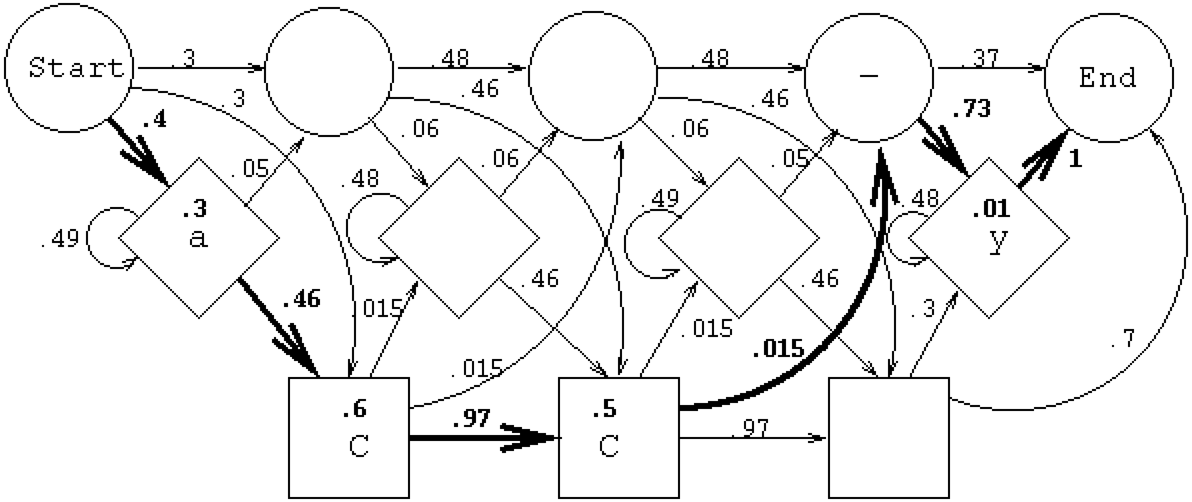


Pyruvate kinase

http://en.wikipedia.org/wiki/File:1pkn.png

# HMM

- They capture what is ‘essential’ to define the family



http://compbio.soe.ucsc.edu/ismb99.handouts/KK185FP.html

# Audience 2

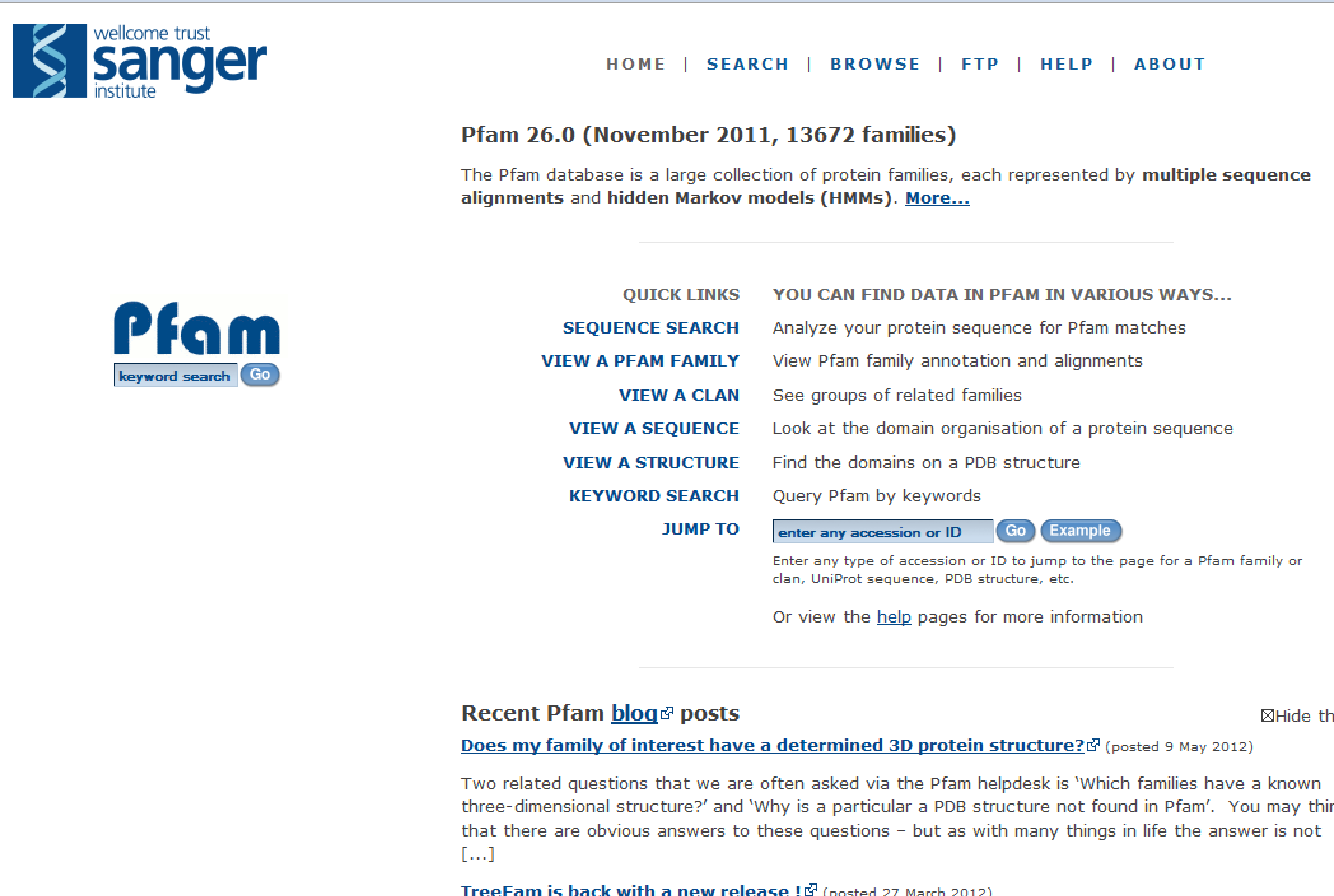
You want to explore and use existing protein family resources for families you are interested in

# Protein family resources

UniProt
Universal Protein Resource
prosite
Pfam
PRINTS
Protein Fingerprints Database
ProDom
SMART
TIGR
THE INSTITUTE FOR GENOMIC RESEARCH
tigr fams
HAMAP
PANTHER
Classification System
PIRSF
Superfamily
Gene3D
Domain Architecture Classification
PDBe
PROTEIN DATA BANK EUROPE
CATH
Protein Structure Classification
SCOP
SWISS-MODEL
ModBase

Clusters of orthologous groups (COG, KOG, eggNOG)

KEGG orthologs



HOME | SEARCH | BROWSE | FTP | HELP | ABOUT

## Pfam 26.0 (November 2011, 13672 families)

The Pfam database is a large collection of protein families, each represented by **multiple sequence alignments** and **hidden Markov models (HMMs)**. More...

| QUICK LINKS | YOU CAN FIND DATA IN PFAM IN VARIOUS WAYS... |
|---|---|
| SEQUENCE SEARCH | Analyze your protein sequence for Pfam matches |
| VIEW A PFAM FAMILY | View Pfam family annotation and alignments |
| VIEW A CLAN | See groups of related families |
| VIEW A SEQUENCE | Look at the domain organisation of a protein sequence |
| VIEW A STRUCTURE | Find the domains on a PDB structure |
| KEYWORD SEARCH | Query Pfam by keywords |
| JUMP TO | enter any accession or ID Go Example |

Enter any type of accession or ID to jump to the page for a Pfam family or clan, UniProt sequence, PDB structure, etc.

Or view the help pages for more information

## Recent Pfam blog posts

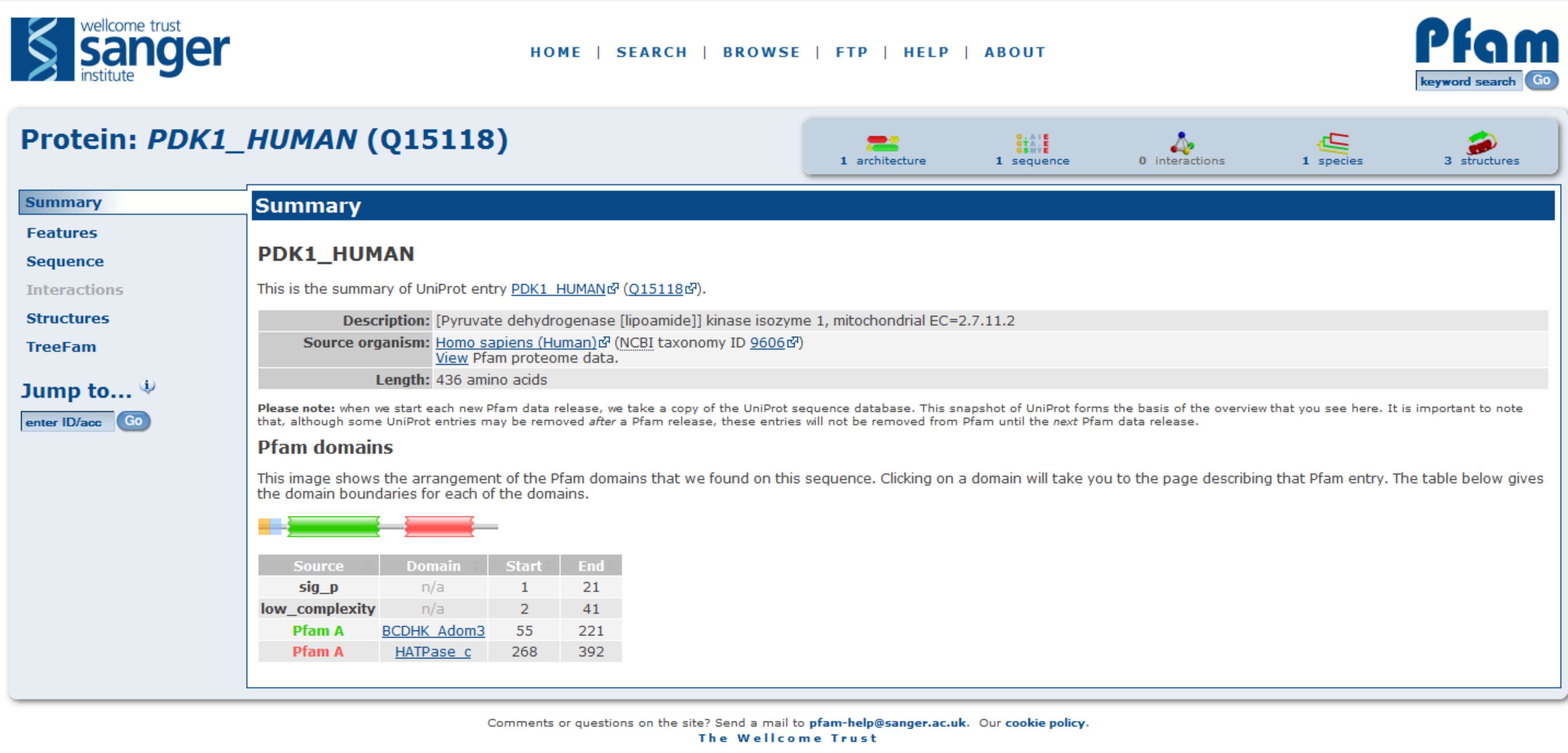
☒Hide th

Does my family of interest have a determined 3D protein structure? (posted 9 May 2012)

Two related questions that we are often asked via the Pfam helpdesk is 'Which families have a known three-dimensional structure?' and 'Why is a particular a PDB structure not found in Pfam'. You may thir that there are obvious answers to these questions – but as with many things in life the answer is not [...]

TreeFam is back with a new release ! (posted 27 March 2012)

# Query by accession



## Protein: *PDK1_HUMAN* (Q15118)

## Summary

### PDK1_HUMAN

This is the summary of UniProt entry PDK1_HUMAN (Q15118).

| | |
|---|---|
| **Description:** | [Pyruvate dehydrogenase [lipoamide]] kinase isozyme 1, mitochondrial EC=2.7.11.2 |
| **Source organism:** | Homo sapiens (Human) (NCBI taxonomy ID 9606)<br>View Pfam proteome data. |
| **Length:** | 436 amino acids |

**Please note:** when we start each new Pfam data release, we take a copy of the UniProt sequence database. This snapshot of UniProt forms the basis of the overview that you see here. It is important to note that, although some UniProt entries may be removed *after* a Pfam release, these entries will not be removed from Pfam until the *next* Pfam data release.

### Pfam domains

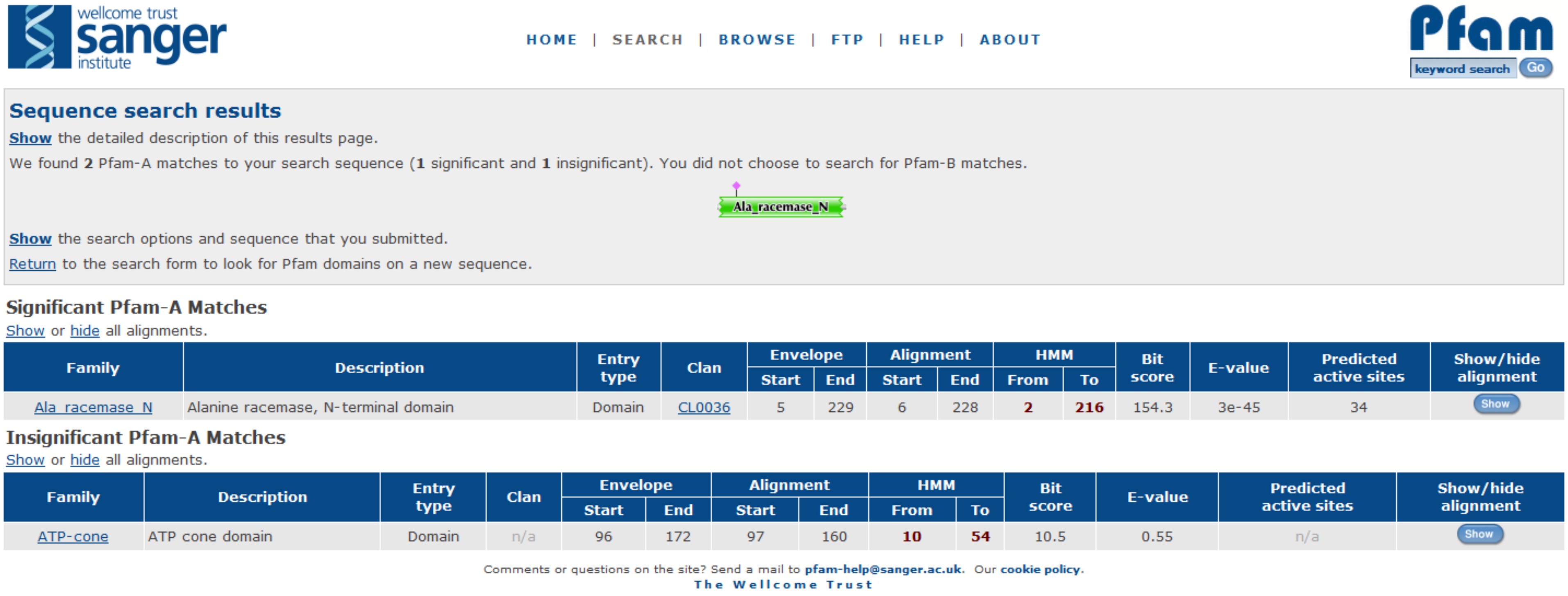
This image shows the arrangement of the Pfam domains that we found on this sequence. Clicking on a domain will take you to the page describing that Pfam entry. The table below gives the domain boundaries for each of the domains.

| Source | Domain | Start | End |
|---|---|---|---|
| sig_p | n/a | 1 | 21 |
| low_complexity | n/a | 2 | 41 |
| Pfam A | BCDHK_Adom3 | 55 | 221 |
| Pfam A | HATPase_c | 268 | 392 |

Comments or questions on the site? Send a mail to **pfam-help@sanger.ac.uk**. Our **cookie policy**.

The Wellcome Trust

# Query by sequence search



wellcome trust sanger institute
HOME | SEARCH | BROWSE | FTP | HELP | ABOUT
Pfam
keyword search
Go
Ala_racemase_N
Show
Show

## Sequence search results

**Show** the detailed description of this results page.

We found **2** Pfam-A matches to your search sequence (**1** significant and **1** insignificant). You did not choose to search for Pfam-B matches.

**Show** the search options and sequence that you submitted.

Return to the search form to look for Pfam domains on a new sequence.

### Significant Pfam-A Matches

Show or hide all alignments.

| Family | Description | Entry type | Clan | Envelope | | Alignment | | HMM | | Bit score | E-value | Predicted active sites | Show/hide alignment |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | Start | End | Start | End | From | To | | | | |
| Ala_racemase_N | Alanine racemase, N-terminal domain | Domain | CL0036 | 5 | 229 | 6 | 228 | **2** | **216** | 154.3 | 3e-45 | 34 | Show |

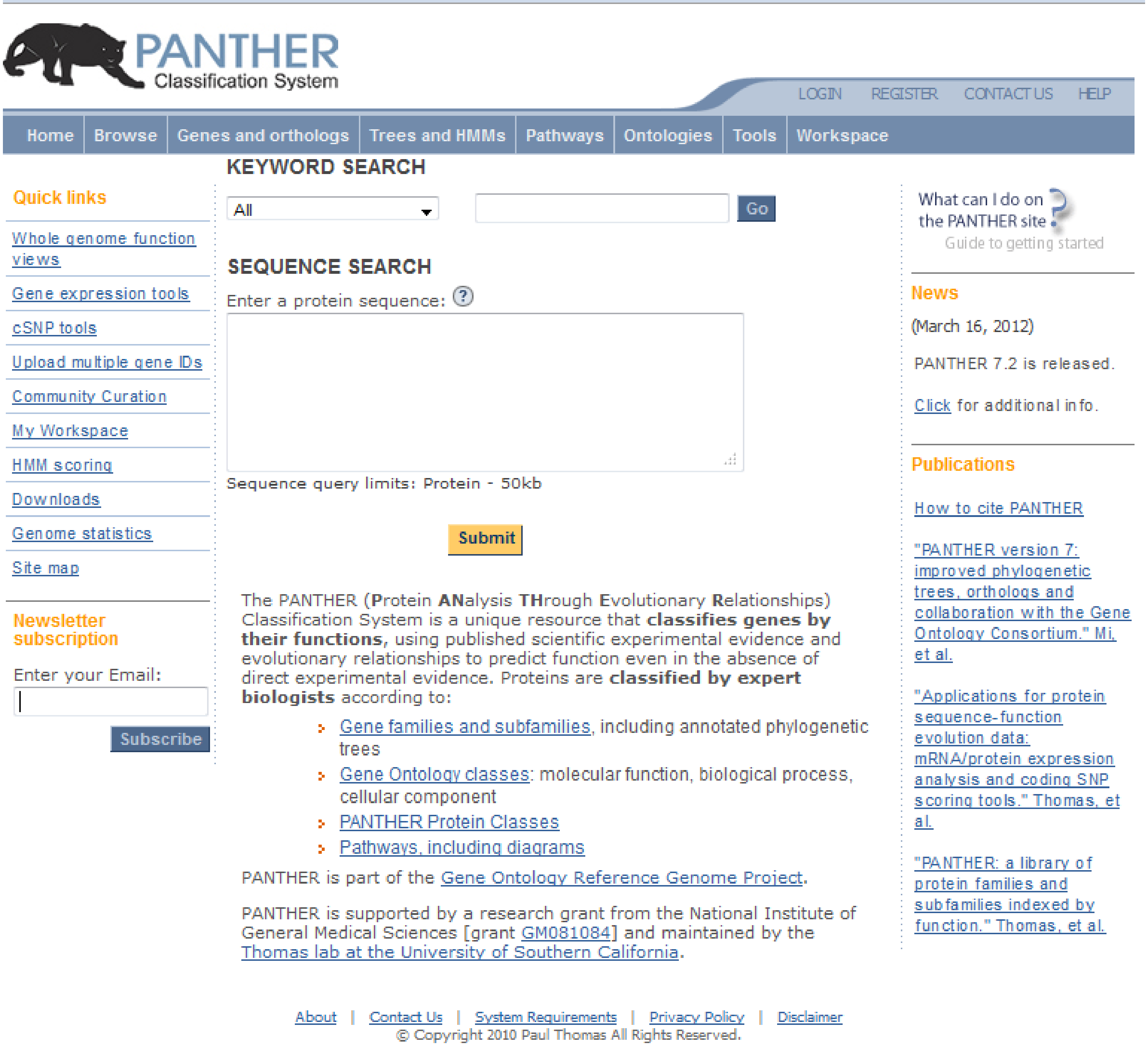
### Insignificant Pfam-A Matches

Show or hide all alignments.

| Family | Description | Entry type | Clan | Envelope | | Alignment | | HMM | | Bit score | E-value | Predicted active sites | Show/hide alignment |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | Start | End | Start | End | From | To | | | | |
| ATP-cone | ATP cone domain | Domain | n/a | 96 | 172 | 97 | 160 | **10** | **54** | 10.5 | 0.55 | n/a | Show |

Comments or questions on the site? Send a mail to **pfam-help@sanger.ac.uk**. Our **cookie policy**.

**The Wellcome Trust**



KEYWORD SEARCH

Quick links

- Whole genome function views
- Gene expression tools
- cSNP tools
- Upload multiple gene IDs
- Community Curation
- My Workspace
- HMM scoring
- Downloads
- Genome statistics
- Site map

Newsletter subscription

Enter your Email:

SEQUENCE SEARCH

Enter a protein sequence: ?

Sequence query limits: Protein - 50kb

The PANTHER (**P**rotein **AN**alysis **TH**rough **E**volutionary **R**elationships) Classification System is a unique resource that **classifies genes by their functions**, using published scientific experimental evidence and evolutionary relationships to predict function even in the absence of direct experimental evidence. Proteins are **classified by expert biologists** according to:

- Gene families and subfamilies, including annotated phylogenetic trees
- Gene Ontology classes: molecular function, biological process, cellular component
- PANTHER Protein Classes
- Pathways, including diagrams

PANTHER is part of the Gene Ontology Reference Genome Project.

PANTHER is supported by a research grant from the National Institute of General Medical Sciences [grant GM081084] and maintained by the Thomas lab at the University of Southern California.

What can I do on the PANTHER site ?

Guide to getting started

News

(March 16, 2012)

PANTHER 7.2 is released.

Click for additional info.

Publications

How to cite PANTHER

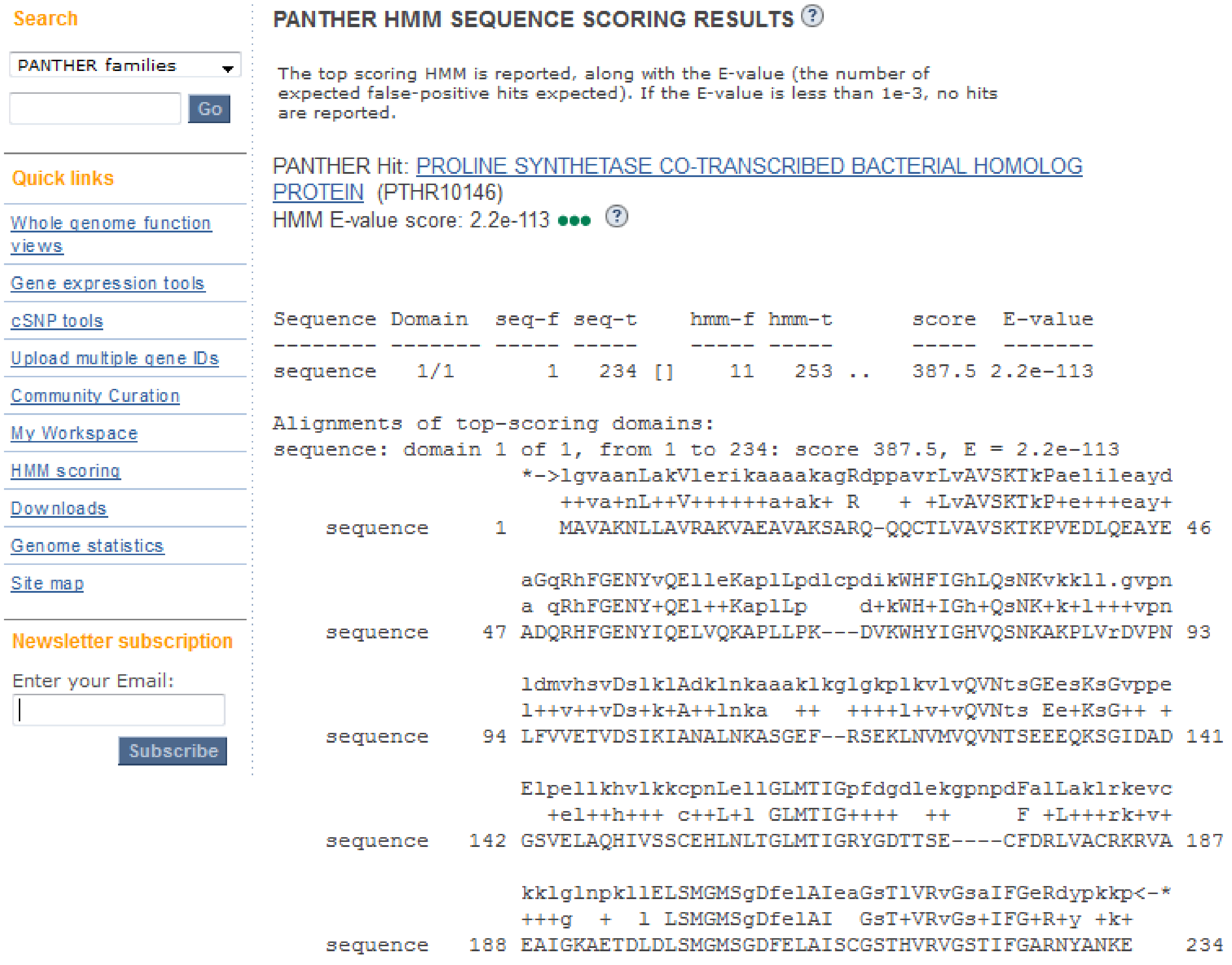
"PANTHER version 7: improved phylogenetic trees, orthologs and collaboration with the Gene Ontology Consortium." Mi, et al.

"Applications for protein sequence-function evolution data: mRNA/protein expression analysis and coding SNP scoring tools." Thomas, et al.

"PANTHER: a library of protein families and subfamilies indexed by function." Thomas, et al.

About | Contact Us | System Requirements | Privacy Policy | Disclaimer

© Copyright 2010 Paul Thomas All Rights Reserved.

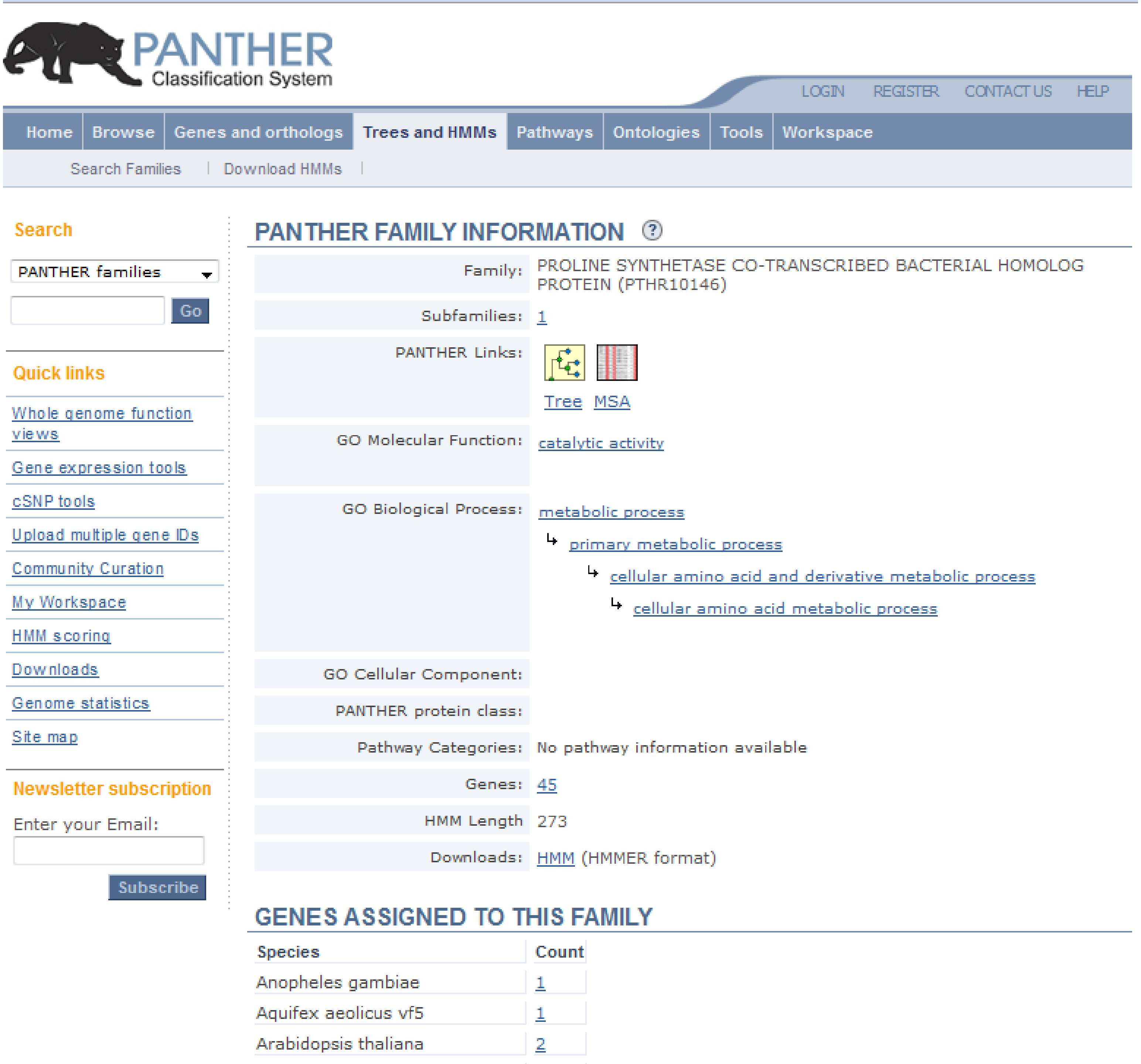


## PANTHER HMM SEQUENCE SCORING RESULTS ?

The top scoring HMM is reported, along with the E-value (the number of expected false-positive hits expected). If the E-value is less than 1e-3, no hits are reported.

PANTHER Hit: PROLINE SYNTHETASE CO-TRANSCRIBED BACTERIAL HOMOLOG PROTEIN (PTHR10146)
HMM E-value score: 2.2e-113 ●●● ?

```
Sequence Domain  seq-f seq-t    hmm-f hmm-t      score  E-value
-------- ------- ----- -----    ----- -----      -----  -------
sequence   1/1       1   234 []    11   253 ..   387.5 2.2e-113

Alignments of top-scoring domains:
sequence: domain 1 of 1, from 1 to 234: score 387.5, E = 2.2e-113
                   *->lgvaanLakVlerikaaaakagRdppavrLvAVSKTkPaelileayd
                      ++va+nL++V++++++a+ak+ R   + +LvAVSKTkP+e+++eay+
    sequence     1    MAVAKNLLAVRAKVAEAVAKSARQ-QQCTLVAVSKTKPVEDLQEAYE 46

                   aGqRhFGENYvQElleKaplLpdlcpdikWHFIGhLQsNKvkkll.gvpn
                   a qRhFGENY+QEl++KaplLp    d+kWH+IGh+QsNK+k+l+++vpn
    sequence    47 ADQRHFGENYIQELVQKAPLLPK---DVKWHYIGHVQSNKAKPLVrDVPN 93

                   ldmvhsvDslklAdklnkaaaklkglgkplkvlvQVNtsGEesKsGvppe
                   l++v++vDs+k+A++lnka  ++  ++++l+v+vQVNts Ee+KsG++ +
    sequence    94 LFVVETVDSIKIANALNKASGEF--RSEKLNVMVQVNTSEEEQKSGIDAD 141

                   ElpellkhvlkkcpnLellGLMTIGpfdgdlekgpnpdFalLaklrkevc
                     +el++h+++ c++L+l GLMTIG++++  ++     F +L+++rk+v+
    sequence   142 GSVELAQHIVSSCEHLNLTGLMTIGRYGDTTSE----CFDRLVACRKRVA 187

                   kklglnpkllELSMGMSgDfelAIeaGsTlVRvGsaIFGeRdypkkp<-*
                   +++g  +  l LSMGMSgDfelAI  GsT+VRvGs+IFG+R+y +k+
    sequence   188 EAIGKAETDLDLSMGMSGDFELAISCGSTHVRVGSTIFGARNYANKE    234
```



## PANTHER FAMILY INFORMATION ?

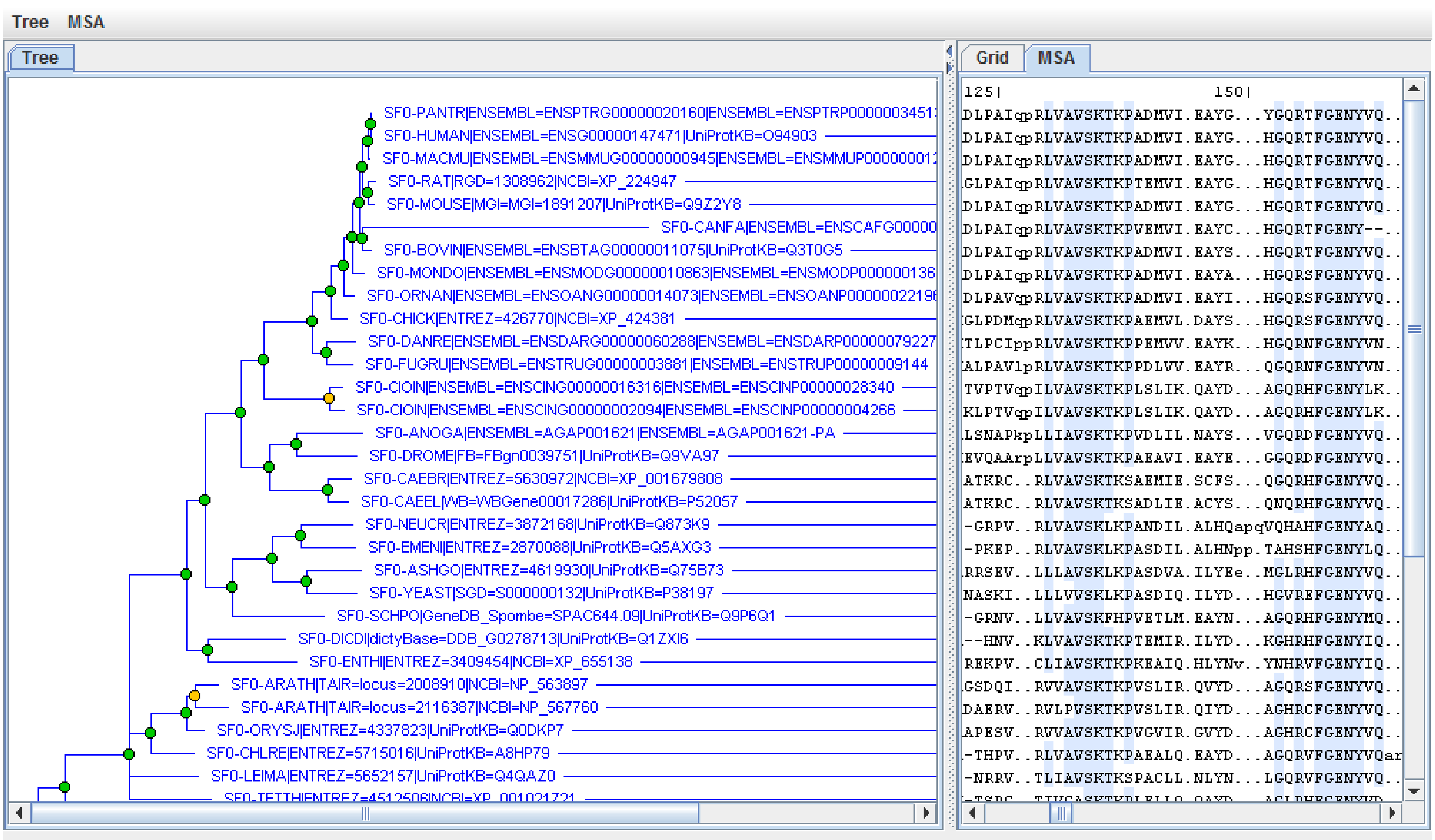
| | |
|---|---|
| Family: | PROLINE SYNTHETASE CO-TRANSCRIBED BACTERIAL HOMOLOG PROTEIN (PTHR10146) |
| Subfamilies: | 1 |
| PANTHER Links: | Tree MSA |
| GO Molecular Function: | catalytic activity |
| GO Biological Process: | metabolic process<br>↳ primary metabolic process<br>↳ cellular amino acid and derivative metabolic process<br>↳ cellular amino acid metabolic process |
| GO Cellular Component: | |
| PANTHER protein class: | |
| Pathway Categories: | No pathway information available |
| Genes: | 45 |
| HMM Length | 273 |
| Downloads: | HMM (HMMER format) |

## GENES ASSIGNED TO THIS FAMILY

| Species | Count |
|---|---|
| Anopheles gambiae | 1 |
| Aquifex aeolicus vf5 | 1 |
| Arabidopsis thaliana | 2 |

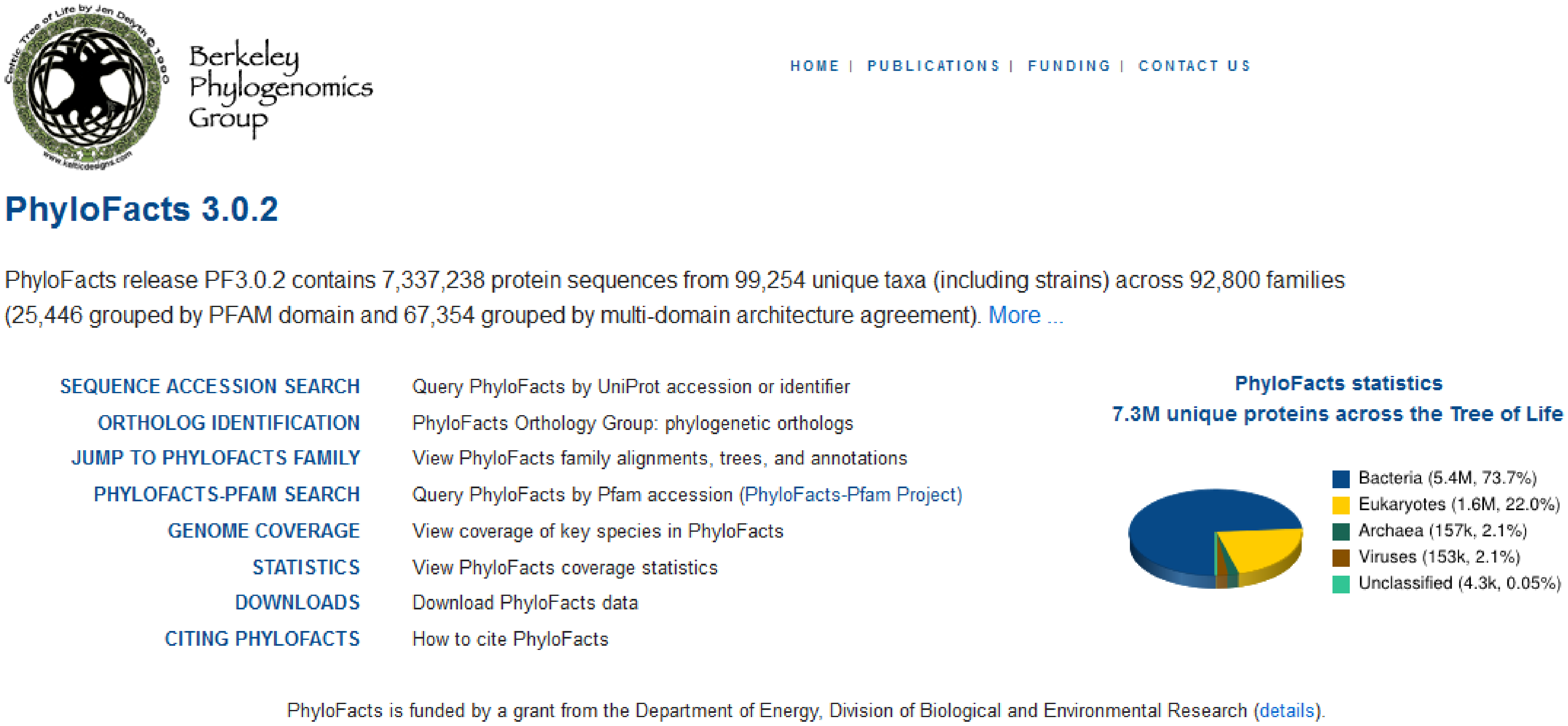
PANTHER TREE VIEWER • close window

Family Name : PROLINE SYNTHETASE CO-TRANSCRIBED BACTERIAL HOMOLOG PROTEIN (PTHR10146)



```
SF0-PANTR|ENSEMBL=ENSPTRG00000020160|ENSEMBL=ENSPTRP0000003451
SF0-HUMAN|ENSEMBL=ENSG00000147471|UniProtKB=O94903
SF0-MACMU|ENSEMBL=ENSMMUG00000000945|ENSEMBL=ENSMMUP00000001
SF0-RAT|RGD=1308962|NCBI=XP_224947
SF0-MOUSE|MGI=MGI=1891207|UniProtKB=Q9Z2Y8
SF0-CANFA|ENSEMBL=ENSCAFG00000
SF0-BOVIN|ENSEMBL=ENSBTAG00000011075|UniProtKB=Q3T0G5
SF0-MONDO|ENSEMBL=ENSMODG00000010863|ENSEMBL=ENSMODP000000136
SF0-ORNAN|ENSEMBL=ENSOANG00000014073|ENSEMBL=ENSOANP00000022196
SF0-CHICK|ENTREZ=426770|NCBI=XP_424381
SF0-DANRE|ENSEMBL=ENSDARG00000060288|ENSEMBL=ENSDARP00000079227
SF0-FUGRU|ENSEMBL=ENSTRUG00000003881|ENSEMBL=ENSTRUP00000009144
SF0-CIOIN|ENSEMBL=ENSCING00000016316|ENSEMBL=ENSCINP00000028340
SF0-CIOIN|ENSEMBL=ENSCING00000002094|ENSEMBL=ENSCINP00000004266
SF0-ANOGA|ENSEMBL=AGAP001621|ENSEMBL=AGAP001621-PA
SF0-DROME|FB=FBgn0039751|UniProtKB=Q9VA97
SF0-CAEBR|ENTREZ=5630972|NCBI=XP_001679808
SF0-CAEEL|WB=WBGene00017286|UniProtKB=P52057
SF0-NEUCR|ENTREZ=3872168|UniProtKB=Q873K9
SF0-EMENI|ENTREZ=2870088|UniProtKB=Q5AXG3
SF0-ASHGO|ENTREZ=4619930|UniProtKB=Q75B73
SF0-YEAST|SGD=S000000132|UniProtKB=P38197
SF0-SCHPO|GeneDB_Spombe=SPAC644.09|UniProtKB=Q9P6Q1
SF0-DICDI|dictyBase=DDB_G0278713|UniProtKB=Q1ZXI6
SF0-ENTHI|ENTREZ=3409454|NCBI=XP_655138
SF0-ARATH|TAIR=locus=2008910|NCBI=NP_563897
SF0-ARATH|TAIR=locus=2116387|NCBI=NP_567760
SF0-ORYSJ|ENTREZ=4337823|UniProtKB=Q0DKP7
SF0-CHLRE|ENTREZ=5715016|UniProtKB=A8HP79
SF0-LEIMA|ENTREZ=5652157|UniProtKB=Q4QAZ0
SF0-TETTH|ENTREZ=4512506|NCBI=XP_001021721
```

```
125|                    150|
DLPAIqpRLVAVSKTKPADMVI.EAYG...YGQRTFGENYVQ..
DLPAIqpRLVAVSKTKPADMVI.EAYG...HGQRTFGENYVQ..
DLPAIqpRLVAVSKTKPADMVI.EAYG...HGQRTFGENYVQ..
GLPAIqpRLVAVSKTKPTEMVI.EAYG...HGQRTFGENYVQ..
DLPAIqpRLVAVSKTKPADMVI.EAYG...HGQRTFGENYVQ..
DLPAIqpRLVAVSKTKPVEMVI.EAYC...HGQRTFGENY--..
DLPAIqpRLVAVSKTKPADMVI.EAYS...HGQRTFGENYVQ..
DLPAIqpRLVAVSKTKPADMVI.EAYA...HGQRSFGENYVQ..
DLPAVqpRLVAVSKTKPADMVI.EAYI...HGQRSFGENYVQ..
GLPDMqpRLVAVSKTKPAEMVL.DAYS...HGQRSFGENYVQ..
TLPCIppRLVAVSKTKPPEMVV.EAYK...HGQRNFGENYVN..
ALPAVlpRLVAVSKTKPPDLVV.EAYR...QGQRNFGENYVN..
TVPTVqpILVAVSKTKPLSLIK.QAYD...AGQRHFGENYLK..
KLPTVqpILVAVSKTKPLSLIK.QAYD...AGQRHFGENYLK..
LSNAPkpLLIAVSKTKPVDLIL.NAYS...VGQRDFGENYVQ..
EVQAArpLLVAVSKTKPAEAVI.EAYE...GGQRDFGENYVQ..
ATKRC..RLVAVSKTKSAEMIE.SCFS...QGQRHFGENYVQ..
ATKRC..RLVAVSKTKSADLIE.ACYS...QNQRHFGENYVQ..
-GRPV..RLVAVSKLKPANDIL.ALHQapqVQHAHFGENYAQ..
-PKEP..RLVAVSKLKPASDIL.ALHNpp.TAHSHFGENYLQ..
RRSEV..LLLAVSKLKPASDVA.ILYEe..MGLRHFGENYVQ..
NASKI..LLLVVSKLKPASDIQ.ILYD...HGVREFGENYVQ..
-GRNV..LLVAVSKFHPVETLM.EAYN...AGQRHFGENYMQ..
--HNV..KLVAVSKTKPTEMIR.ILYD...KGHRHFGENYIQ..
REKPV..CLIAVSKTKPKEAIQ.HLYNv..YNHRVFGENYIQ..
GSDQI..RVVAVSKTKPVSLIR.QVYD...AGQRSFGENYVQ..
DAERV..RVLPVSKTKPVSLIR.QIYD...AGHRCFGENYVQ..
APESV..RVVAVSKTKPVGVIR.GVYD...AGHRCFGENYVQ..
-THPV..RLVAVSKTKPAEALQ.EAYD...AGQRVFGENYVQar
-NRRV..TLIAVSKTKSPACLL.NLYN...LGQRVFGENYVQ..
-TSRC..TIVCASKTKPLELLQ.QAYD...ACLRHFGENYVD
```



Berkeley
Phylogenomics
Group

HOME | PUBLICATIONS | FUNDING | CONTACT US

# PhyloFacts 3.0.2

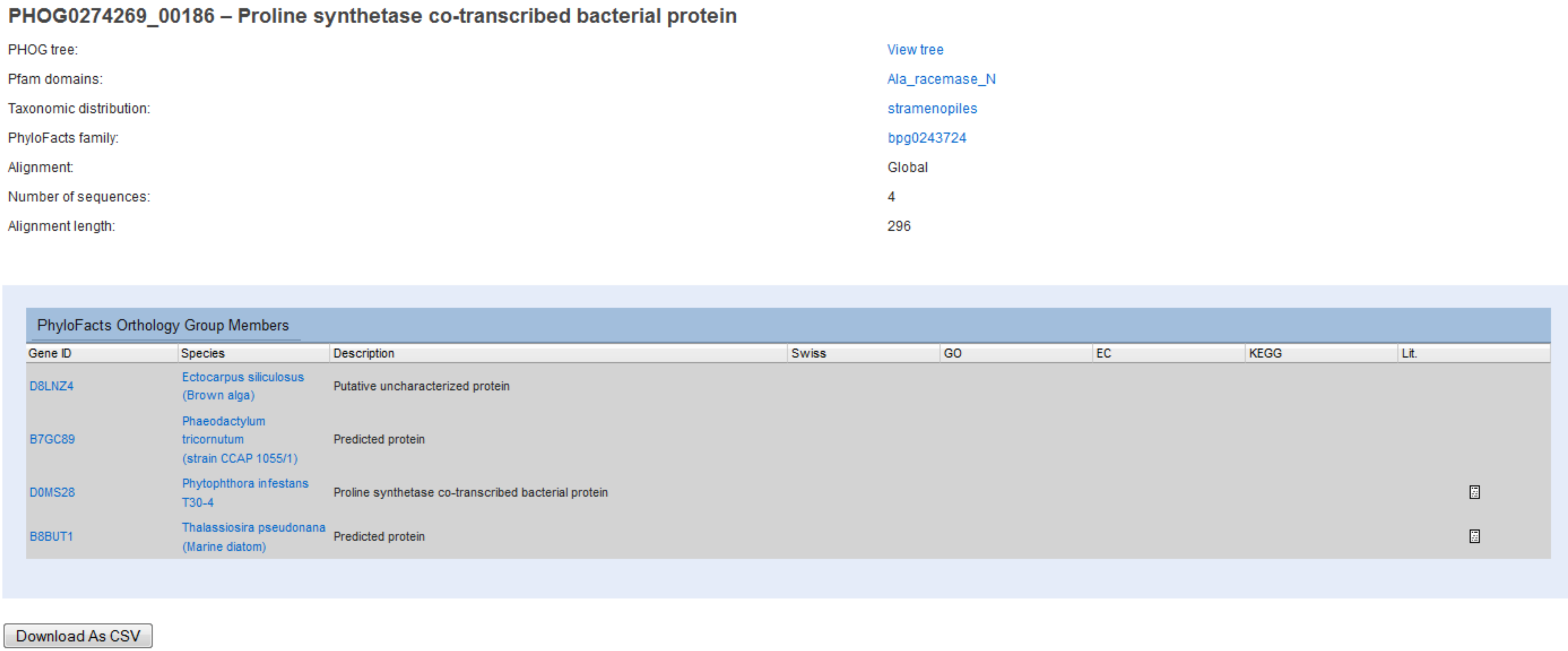
PhyloFacts release PF3.0.2 contains 7,337,238 protein sequences from 99,254 unique taxa (including strains) across 92,800 families (25,446 grouped by PFAM domain and 67,354 grouped by multi-domain architecture agreement). More ...

| | |
|---|---|
| SEQUENCE ACCESSION SEARCH | Query PhyloFacts by UniProt accession or identifier |
| ORTHOLOG IDENTIFICATION | PhyloFacts Orthology Group: phylogenetic orthologs |
| JUMP TO PHYLOFACTS FAMILY | View PhyloFacts family alignments, trees, and annotations |
| PHYLOFACTS-PFAM SEARCH | Query PhyloFacts by Pfam accession (PhyloFacts-Pfam Project) |
| GENOME COVERAGE | View coverage of key species in PhyloFacts |
| STATISTICS | View PhyloFacts coverage statistics |
| DOWNLOADS | Download PhyloFacts data |
| CITING PHYLOFACTS | How to cite PhyloFacts |

PhyloFacts statistics
7.3M unique proteins across the Tree of Life

PhyloFacts is funded by a grant from the Department of Energy, Division of Biological and Environmental Research (details).

# Phylofacts
# query by sequence search



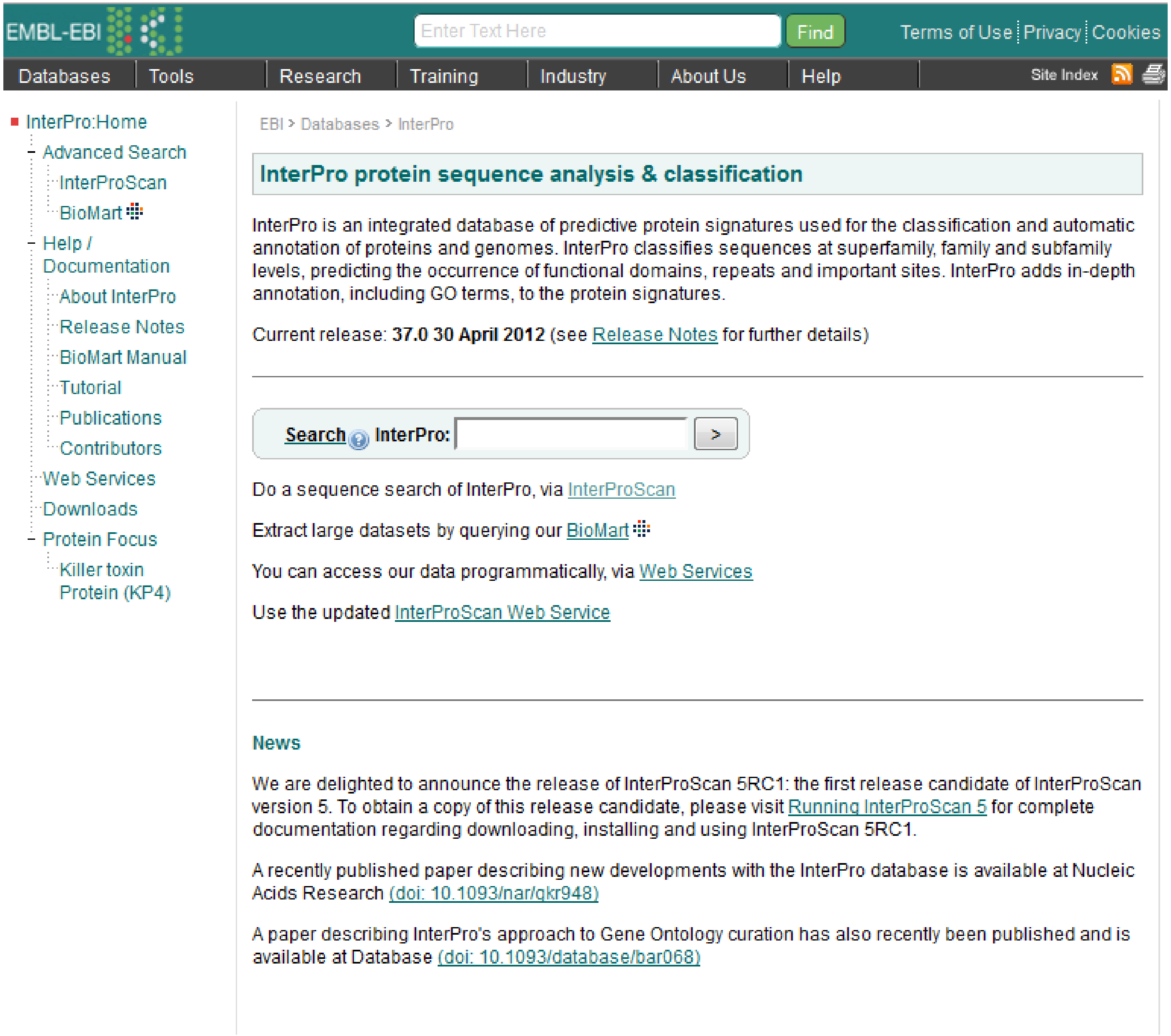
### PHOG0274269_00186 – Proline synthetase co-transcribed bacterial protein

| | |
|---|---|
| PHOG tree: | View tree |
| Pfam domains: | Ala_racemase_N |
| Taxonomic distribution: | stramenopiles |
| PhyloFacts family: | bpg0243724 |
| Alignment: | Global |
| Number of sequences: | 4 |
| Alignment length: | 296 |

PhyloFacts Orthology Group Members

| Gene ID | Species | Description | Swiss | GO | EC | KEGG | Lit. |
|---|---|---|---|---|---|---|---|
| D8LNZ4 | Ectocarpus siliculosus (Brown alga) | Putative uncharacterized protein | | | | | |
| B7GC89 | Phaeodactylum tricornutum (strain CCAP 1055/1) | Predicted protein | | | | | |
| D0MS28 | Phytophthora infestans T30-4 | Proline synthetase co-transcribed bacterial protein | | | | | ▤ |
| B8BUT1 | Thalassiosira pseudonana (Marine diatom) | Predicted protein | | | | | ▤ |

# Resource federation: InterPro



- InterPro:Home
  - Advanced Search
    - InterProScan
    - BioMart
  - Help / Documentation
    - About InterPro
    - Release Notes
    - BioMart Manual
    - Tutorial
    - Publications
    - Contributors
  - Web Services
  - Downloads
  - Protein Focus
    - Killer toxin Protein (KP4)

EBI > Databases > InterPro

## InterPro protein sequence analysis & classification

InterPro is an integrated database of predictive protein signatures used for the classification and automatic annotation of proteins and genomes. InterPro classifies sequences at superfamily, family and subfamily levels, predicting the occurrence of functional domains, repeats and important sites. InterPro adds in-depth annotation, including GO terms, to the protein signatures.

Current release: **37.0 30 April 2012** (see Release Notes for further details)

**Search** **InterPro:** >

Do a sequence search of InterPro, via InterProScan

Extract large datasets by querying our BioMart

You can access our data programmatically, via Web Services

Use the updated InterProScan Web Service

### News

We are delighted to announce the release of InterProScan 5RC1: the first release candidate of InterProScan version 5. To obtain a copy of this release candidate, please visit Running InterProScan 5 for complete documentation regarding downloading, installing and using InterProScan 5RC1.

A recently published paper describing new developments with the InterPro database is available at Nucleic Acids Research (doi: 10.1093/nar/gkr948)

A paper describing InterPro's approach to Gene Ontology curation has also recently been published and is available at Database (doi: 10.1093/database/bar068)

# Not as easy as it may sound…

- Specific protein families may not be consistent across resources
- Most families (MSAs, trees, HMMs) in these resources are not manually curated
    - Domains in Pfam-A are curated
    - TIGRfams are curated
    - HAMAP families are curated

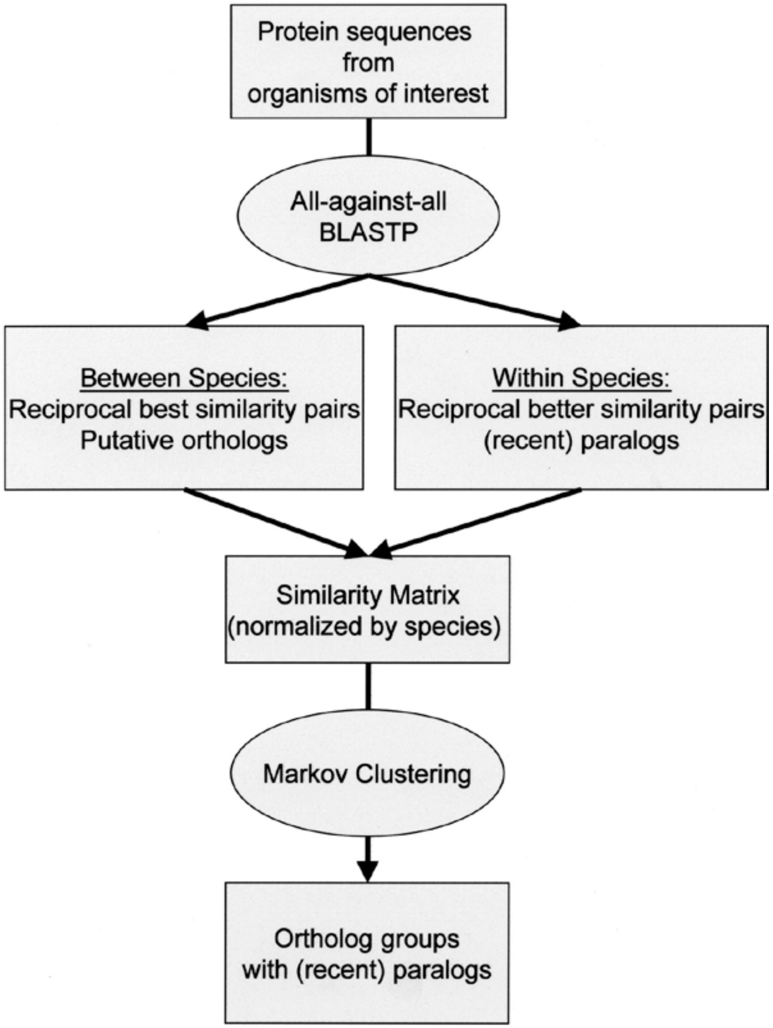
# Audience 3

You have a new genome and you want to place all of its genes into their respective families

# Solutions

- Build one at a time (impractical)
- orthoMCL
- multiParanoid

# orthoMCL pipeline



Protein sequences from organisms of interest

All-against-all BLASTP

Between Species:
Reciprocal best similarity pairs
Putative orthologs

Within Species:
Reciprocal better similarity pairs
(recent) paralogs

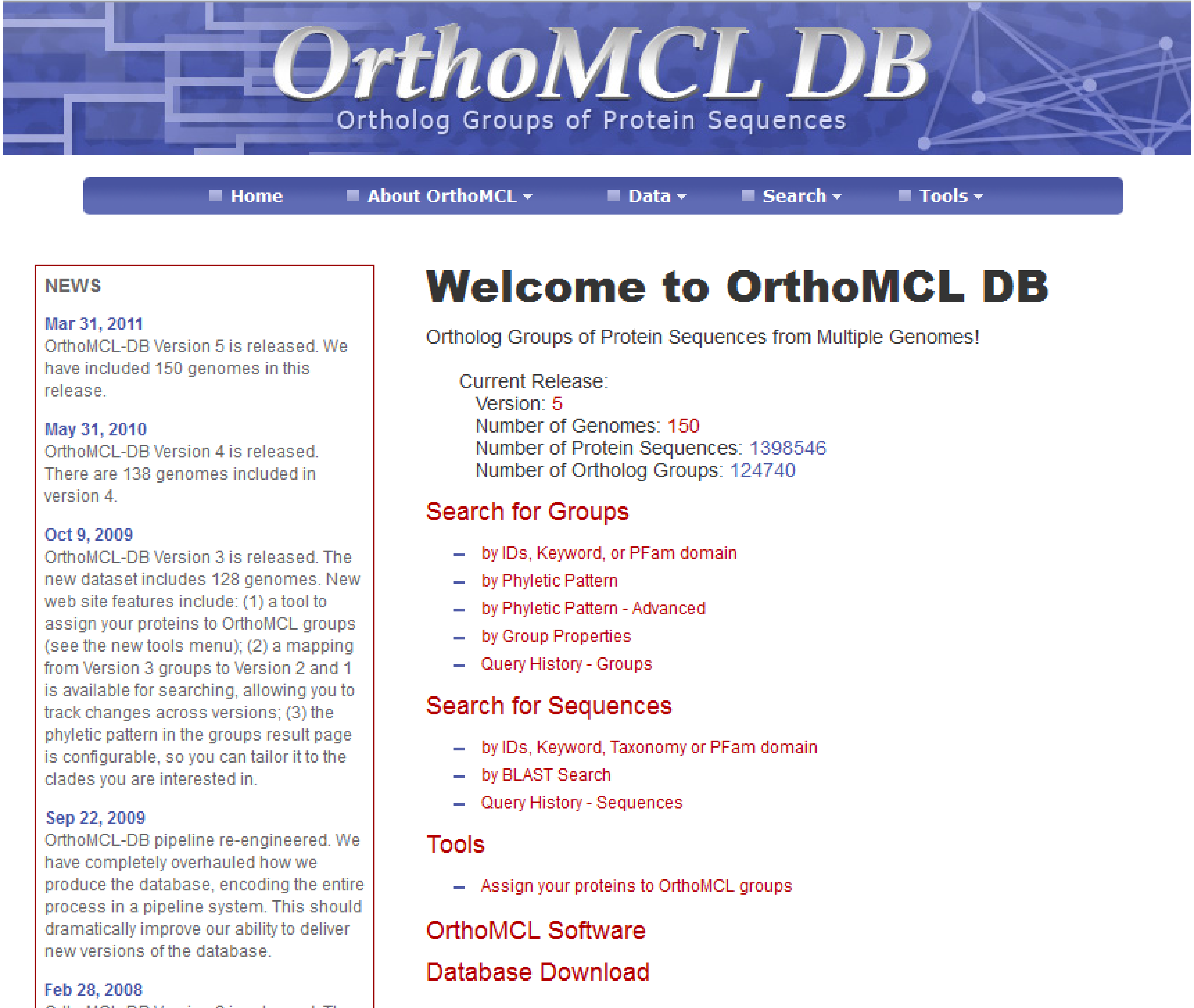
Similarity Matrix
(normalized by species)

Markov Clustering

Ortholog groups
with (recent) paralogs

**Li Li et al. Genome Res. 2003; 13: 2178-2189**

Cold Spring Harbor Laboratory Press



NEWS

**Mar 31, 2011**
OrthoMCL-DB Version 5 is released. We have included 150 genomes in this release.

**May 31, 2010**
OrthoMCL-DB Version 4 is released. There are 138 genomes included in version 4.

**Oct 9, 2009**
OrthoMCL-DB Version 3 is released. The new dataset includes 128 genomes. New web site features include: (1) a tool to assign your proteins to OrthoMCL groups (see the new tools menu); (2) a mapping from Version 3 groups to Version 2 and 1 is available for searching, allowing you to track changes across versions; (3) the phyletic pattern in the groups result page is configurable, so you can tailor it to the clades you are interested in.

**Sep 22, 2009**
OrthoMCL-DB pipeline re-engineered. We have completely overhauled how we produce the database, encoding the entire process in a pipeline system. This should dramatically improve our ability to deliver new versions of the database.

**Feb 28, 2008**

# Welcome to OrthoMCL DB

Ortholog Groups of Protein Sequences from Multiple Genomes!

Current Release:
Version: 5
Number of Genomes: 150
Number of Protein Sequences: 1398546
Number of Ortholog Groups: 124740

## Search for Groups

- by IDs, Keyword, or PFam domain
- by Phyletic Pattern
- by Phyletic Pattern - Advanced
- by Group Properties
- Query History - Groups

## Search for Sequences

- by IDs, Keyword, Taxonomy or PFam domain
- by BLAST Search
- Query History - Sequences

## Tools

- Assign your proteins to OrthoMCL groups

## OrthoMCL Software

## Database Download

# Final remarks

- The need for experimental results
- Conserved hypotheticals
  - The domino effect