

Functional Annotation of proteins

Johann Joets

Institut Diversité, Evolution, Ecologie du Vivant

INRAE – Université Paris-Saclay

École "Assemblage & Annotation" AVIESAN 2022 - Roscoff

Functional annotation of genes/proteins



[SEARCH](#) [BLAST](#) [BROWSE](#) [ANNOTATIONS](#) [MCL CLUSTERS](#) [SYNTENY](#) [DOWNLOAD](#) [INFO](#) [HOME](#) [STATUS](#) [HELP!](#)

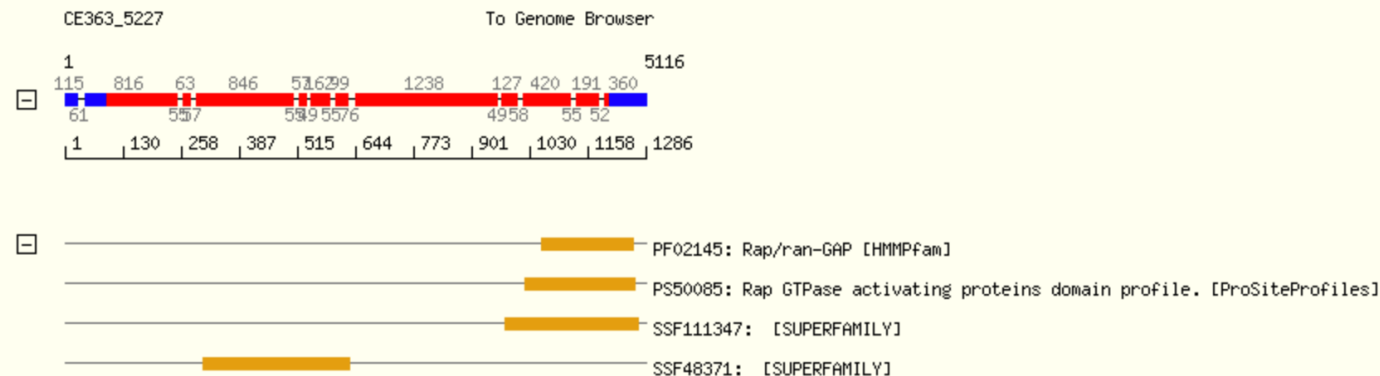
Name: CE363_5227
 Protein ID: 364
 Location: scaffold_1:102962-108077
 Strand: +
 Number of exons: 12
 Description: Longest ORF from: 314 to 4171 breakup#1
 Best Hit: [gij384493585|gb|E|E84076.1|_hypothetical protein RO3G_08781 \[Rhizopus oryzae RA 99-880\].\(model%: 89, hit%: 93, score: 3302, %id: 58\) \[Rhizopus delemar RA 99-880\]](#)
 total hits(shown) 360 (0)

ASPECT	GO Id	GO Desc	Interpro Id	Interpro Desc
Molecular Function	0005096	GTPase activator activity	IPR000331	Rap GTPase activating protein domain
	0005488	binding	IPR016024	Armadillo-type fold
KOG GROUP	KOG Id	KOG Class		KOG Desc
Cellular Processes And Signaling	KOG3686	Signal transduction mechanisms		Rap1-GTPase-activating protein (Rap1GAP)

[View/modify manual annotation](#)

[View nucleotide and 3-frame translation](#) [To Genome Browser](#)

[NCBI blastp](#) Predicted number of transmembrane domains: 0



Functional annotation of genes/proteins



MycoCosm

THE FUNGAL GENOMICS RESOURCE

[JGI HOME](#) [GENOME PORTAL](#) [MYCOCOSM](#) [PHYCOCOSM](#) [LOGOUT](#) [JOHANN JOETS \(JOHANN.JOETS@INRAE.FR\)](#)

Browse • *Mucor mucedo* NRRL 3635 v1.0

[SEARCH](#) [BLAST](#) [BROWSE](#) [ANNOTATIONS](#) [MCL CLUSTERS](#) [SYNTENY](#) [DOWNLOAD](#) [INFO](#) [HOME](#) [STATUS](#) [HELP!](#)

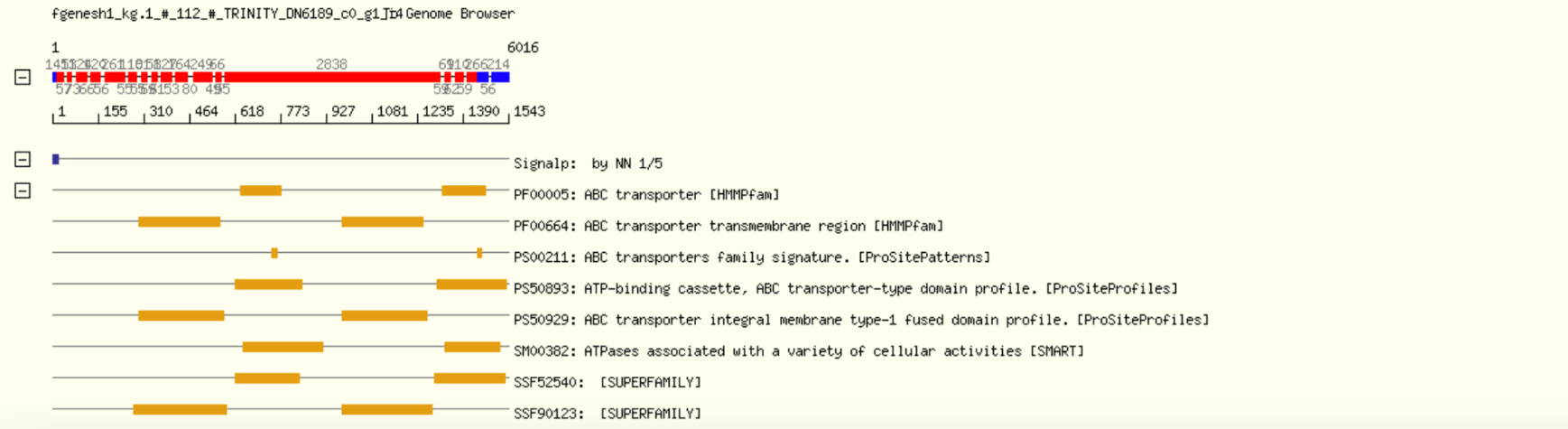
Name: fgenes1_kg.1_#_112_#_TRINITY_DN6189_c0_g1_i4
 Protein ID: 489278
 Location: scaffold_1:163512-169527
 Strand: -
 Number of exons: 17
 Description:
 Best Hit: gj|384495448|gb|EIE85939.1| multi drug resistance-associated protein MRP [Rhizopus oryzae RA 99-880] (model%: 93, hit%: 99, score: 5020, %id: 65) [Rhizopus delemar RA 99-880]
 total hits(shown) 800 (10)

ASPECT	GO Id	GO Desc	Interpro Id	Interpro Desc
Molecular Function	0042626	ATPase activity, coupled to transmembrane movement of substances	IPR011527	ABC transporter type 1, transmembrane domain
	0005524	ATP binding	IPR003439	ABC transporter-like
			IPR011527	ABC transporter type 1, transmembrane domain
			IPR017871	ABC transporter, conserved site
Biological Process	0016887	ATPase activity	IPR003439	ABC transporter-like
	0006810	transport	IPR017871	ABC transporter, conserved site
Cellular Component	0016021	integral to membrane	IPR011527	ABC transporter type 1, transmembrane domain
KOG GROUP	KOG Id	KOG Class	KOG Desc	
Metabolism	KOG0054	Secondary metabolites biosynthesis, transport and catabolism	Multidrug resistance-associated protein/mitoxantrone resistance protein, ABC superfamily	


[View/modify manual annotation](#)

[View nucleotide and 3-frame translation](#) [To Genome Browser](#)

NCBI blastp Predicted number of transmembrane domains: 14



Functional annotation of genes/proteins

 **InterPro** Classification of protein families 🔍 ☰

Home ▶ Search ▶ **Browse** ▶ Results Release notes Download ▶ Help ▶ About



ABC transporter type 1, transmembrane domain [★]

IPR011527

[InterPro entry](#)

◀	
Overview	
Proteins	493k
Domain Architectures	2k
Taxonomy	46k
Proteomes	11k
Structures	223
RoseTTAFold	1
AlphaFold	418k
Pathways	244

Short name: *ABC1_TM_dom*

Overlapping homologous superfamilies ⓘ

[H](#) [ABC transporter type 1, transmembrane domain superfamily](#) (IPR036640)

Domain relationships

- ▼ **D** [ABC transporter type 1, transmembrane domain](#) (IPR011527)
 - D** [ABC transporter C family, six-transmembrane helical domain 2](#) (IPR044726)
 - D** [ABC transporter C family, six-transmembrane helical domain 1](#) (IPR044746)

Description

ABC transporters belong to the ATP-Binding Cassette (ABC) superfamily, which uses the hydrolysis of ATP to energise diverse biological systems. ABC transporters minimally consist of two conserved regions: a highly conserved ATP binding cassette (ABC) and a less conserved transmembrane domain (TMD). These can be found on the same protein or on two different ones. Most ABC transporters function as a dimer and therefore are constituted of four domains, two ABC modules and two TMDs.

GO terms

Biological Process

- transmembrane transport (GO:0055085) [↗](#)

Molecular Function

- ATP binding (GO:0005524) [↗](#)
- ABC-type transporter activity (GO:0140359) [↗](#)

Cellular Component

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

References

1. [^] Getting in or out: early segregation between importers and exporters in the evolution of ATP-binding cassette (ABC) transporters. Saurin W, Hofnung M, Dassa E. *J. Mol. Evol.* 48, 22-41, (1999). [View article](#) [↗](#) PMID: 9873074 [↗](#)

2. [^] The ABC of ABCS: a phylogenetic and functional classification of ABC systems in living organisms. Dassa E, Bouige P. *Res. Microbiol.* 152, 211-29, (2001). [View article](#) [↗](#) PMID: 11421270 [↗](#)

3. [^] ABC transporters: physiology, structure and mechanism--an overview. Higgins CF. *Res. Microbiol.* 152, 205-10, (2001). [View article](#) [↗](#) PMID: 11421269 [↗](#)

4. [^] ABC-ATPases, adaptable energy generators fuelling transmembrane movement of a variety of molecules in organisms from bacteria to humans. Holland IB, Blight MA. *J. Mol. Biol.* 293, 381-99, (1999). [View article](#) [↗](#) PMID: 10529352 [↗](#)

[Add your annotation](#) ▼

Contributing Member Database Entries



PROSITE profiles:
[PS50929](#)



Pfam: [PF13748](#),
[PF06472](#),[PF00664](#)

Functional annotation of genes/proteins



MycoCosm

THE FUNGAL GENOMICS RESOURCE

[JGI HOME](#) [GENOME PORTAL](#) [MYCOCOSM](#) [PHYCOCOSM](#) [LOGOUT](#) [JOHANN JOETS \(JOHANN.JOETS@INRAE.FR\)](#)

Browse • Mucor mucedo NRRL 3635 v1.0

[SEARCH](#) [BLAST](#) [BROWSE](#) [ANNOTATIONS](#) [MCL CLUSTERS](#) [SYNTENY](#) [DOWNLOAD](#) [INFO](#) [HOME](#) [STATUS](#) [HELP!](#)

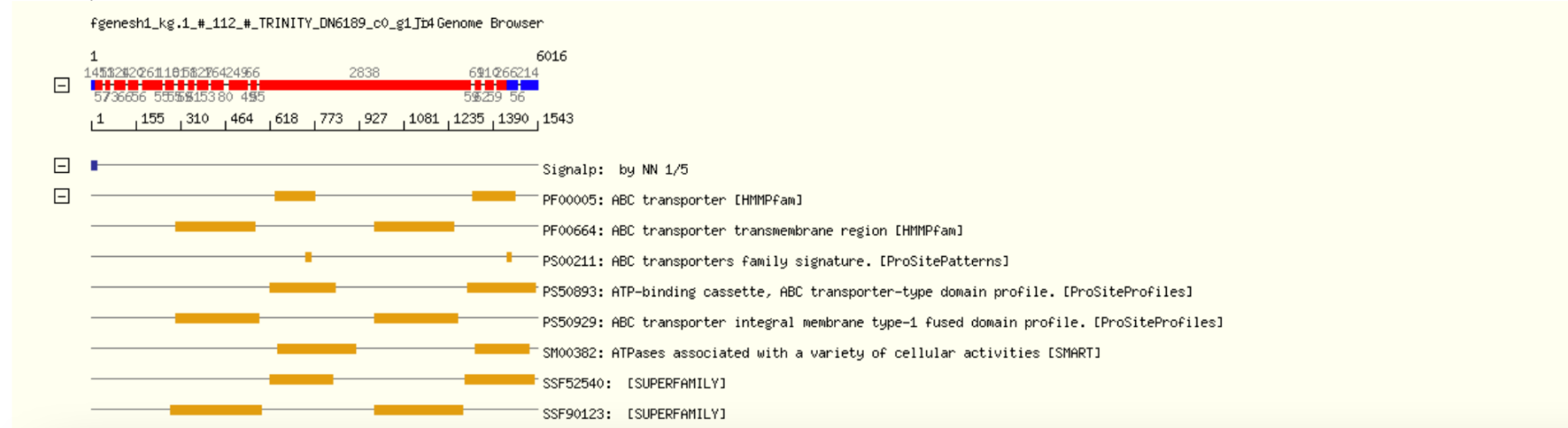
Name: fgenes1_kg.1_#_112_#_TRINITY_DN6189_c0_g1_i4
 Protein ID: 489278
 Location: scaffold_1:163512-169527
 Strand: -
 Number of exons: 17
 Description:
 Best Hit: gjl384495448|gb|EIE85939.1| multi drug resistance-associated protein MRP [Rhizopus oryzae RA 99-880] (model%: 93, hit%: 99, score: 5020, %id: 65) [Rhizopus delemar RA 99-880]
 total hits(shown) 800 (10)

ASPECT	GO Id	GO Desc	Interpro Id	Interpro Desc
Molecular Function	0042626	ATPase activity, coupled to transmembrane movement of substances	IPR011527	ABC transporter type 1, transmembrane domain
	0005524	ATP binding	IPR003439	ABC transporter-like
			IPR011527	ABC transporter type 1, transmembrane domain
			IPR017871	ABC transporter, conserved site
			IPR003439	ABC transporter-like
Biological Process	0016887	ATPase activity	IPR017871	ABC transporter, conserved site
	0006810	transport	IPR011527	ABC transporter type 1, transmembrane domain
Cellular Component	0016021	integral to membrane	IPR011527	ABC transporter type 1, transmembrane domain
KOG GROUP	KOG Id	KOG Class	KOG Desc	
Metabolism	KOG0054	Secondary metabolites biosynthesis, transport and catabolism	Multidrug resistance-associated protein/mitoxantrone resistance protein, ABC superfamily	


[View/modify manual annotation](#)

[View nucleotide and 3-frame translation](#) [To Genome Browser](#)

[NCBI blastp](#) [Predicted number of transmembrane domains: 14](#)










Functional annotation of genes/proteins



Q

Navigation

-  Home
-  Sequence search
-  eggNOG-mapper v2
(Functional annotation)
(based on eggNOG v5.0)
-  Downloads
-  API
-  Methods
-  Viral OGs

Query OG
✕

KOG0054

Q

KOG0054

Eukaryotes

Q Secondary metabolites biosynthesis, transport, and catabolism

ATP-binding cassette, sub-family C (CFTR MRP), member

3380 proteins

233 species

Ortholog	Organism
TRIADP9790	<i>Trichoplax adhaerens</i>
ABCC12	<i>Pelodiscus sinensis</i>
BT.104466	<i>Bos taurus</i>
XP_001451599.1	<i>Paramecium tetraurelia strain d4 2</i>
SCAFFOLD_503549.1	<i>Arabidopsis lyrata</i>
CADACLAP00004315	<i>Aspergillus clavatus</i>
CMU_010280	<i>Cryptosporidium muris RN66</i>
MRPA	<i>Leishmania infantum JPCMS</i>

3372 more...

Fine-grained Orthologs
Orthologous Group
Taxonomic Profile
Functional Profile
Alignment
Phylogenetic Tree
Download ▾

Gene Ontology

KEGG pathways

Domains

Frequency of KEGG pathways

KEGG Pathways		
Pathway	SeqCount	Frequency
ABC transporters (02010)	308	9.11%

Found 1 matches in 0.01 seconds (1KB)

Functional annotation of genes/proteins



MycoCosm

THE FUNGAL GENOMICS RESOURCE

[JGI HOME](#) [GENOME PORTAL](#) [MYCOCOSM](#) [PHYCOCOSM](#) [LOGOUT](#) [JOHANN JOETS \(JOHANN.JOETS@INRAE.FR\)](#)

Browse • Mucor mucedo NRRL 3635 v1.0

[SEARCH](#) [BLAST](#) [BROWSE](#) [ANNOTATIONS](#) [MCL CLUSTERS](#) [SYNTENY](#) [DOWNLOAD](#) [INFO](#) [HOME](#) [STATUS](#) [HELP!](#)

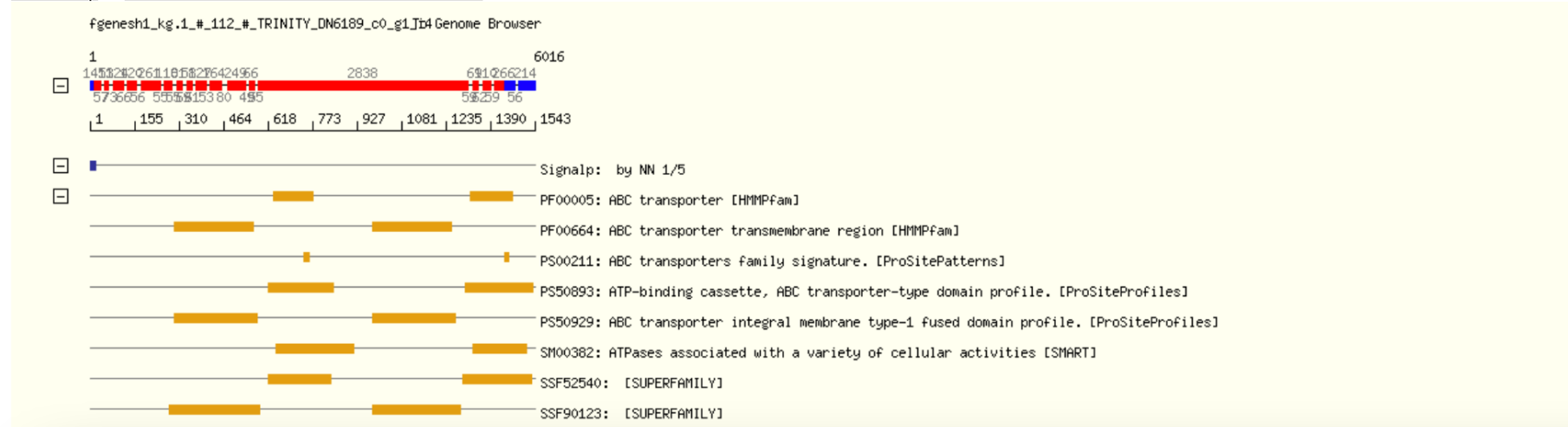
Name: fgenes1_kg.1_#_112_#_TRINITY_DN6189_c0_g1_i4
 Protein ID: 489278
 Location: scaffold_1:163512-169527
 Strand: -
 Number of exons: 17
 Description:
 Best Hit: gj|384495448|gb|EIE85939.1| multi drug resistance-associated protein MRP [Rhizopus oryzae RA 99-880] (model%: 93, hit%: 99, score: 5020, %id: 65) [Rhizopus delemar RA 99-880]
 total hits(shown) 800 (10)

ASPECT	GO Id	GO Desc	Interpro Id	Interpro Desc
Molecular Function	0042626	ATPase activity, coupled to transmembrane movement of substances	IPR011527	ABC transporter type 1, transmembrane domain
	0005524	ATP binding	IPR003439	ABC transporter-like
			IPR011527	ABC transporter type 1, transmembrane domain
			IPR017871	ABC transporter, conserved site
Biological Process	0016887	ATPase activity	IPR003439	ABC transporter-like
	IPR017871	ABC transporter, conserved site		
Cellular Component	0006810	transport	IPR011527	ABC transporter type 1, transmembrane domain
	0016021	integral to membrane	IPR011527	ABC transporter type 1, transmembrane domain
KOG GROUP	KOG Id	KOG Class	KOG Desc	
Metabolism	KOG0054	Secondary metabolites biosynthesis, transport and catabolism	Multidrug resistance-associated protein/mitoxantrone resistance protein, ABC superfamily	

[View/modify manual annotation](#)

[View nucleotide and 3-frame translation](#) [To Genome Browser](#)

NCBI blastp Predicted number of transmembrane domains: 14



Functional annotation of genes/proteins

Quick GO Search

Help Contact API Basket

Overview
Synonyms
Ancestor Chart
Child Terms
Annotation Guidance
GO Discussions
Taxon Constraints
Blacklist
Cross-References
Cross-Ontology Relations
Replaces
Replaced By

GO:0042626 F 🛒 🔗 JSON

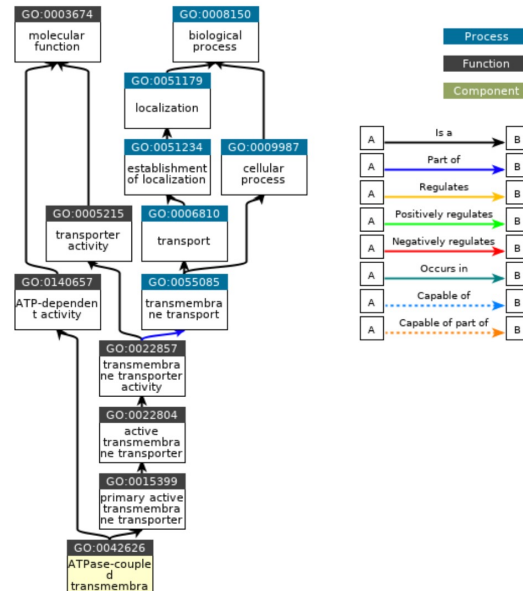
ATPase-coupled transmembrane transporter activity

Molecular Function

Definition ([GO:0042626 GONUTS page](#))
Primary active transporter of a solute across a membrane, via the reaction: $ATP + H_2O = ADP + \text{phosphate}$, to directly drive the transport of a substance across a membrane. The transport protein may be transiently phosphorylated (P-type transporters), or not (ABC-type transporters and other families of transporters). Primary active transport occurs up the solute's concentration gradient and is driven by a primary energy source.

Ancestor Chart 🗨

Ancestor chart for GO:0042626 Chart options ▾



Child Terms

This table lists all terms that are direct descendants (child terms) of GO:0042626

Child Term	Relationship to GO:0042626
GO:1901514 F 🛒 🔗 ATPase-coupled lipo-chitin oligosaccharide transmembrane transporter activity	is_a
GO:0034040 F 🛒 🔗 ATPase-coupled lipid transmembrane transporter activity	is_a
GO:0019829 F 🛒 🔗 ATPase-coupled cation transmembrane transporter activity	is_a
GO:0042625 F 🛒 🔗 ATPase-coupled ion transmembrane transporter activity	is_a
GO:0033225 F 🛒 🔗 ATPase-coupled 2-aminoethylphosphonate transporter activity	is_a
GO:0015450 F 🛒 🔗 protein-transporting ATPase activity	is_a
GO:0033221 F 🛒 🔗 ATPase-coupled urea transmembrane transporter activity	is_a
GO:0098533 C 🛒 🔗 ATPase dependent transmembrane transport complex	capable_of
GO:0043225 F 🛒 🔗 ATPase-coupled inorganic anion transmembrane transporter activity	is_a
GO:0140359 F 🛒 🔗 ABC-type transporter activity	is_a

Functional annotation pipelines



Ten steps to get started in Genome Assembly and Annotation

[version 1; peer review: 2 approved]

Victoria Dominguez Del Angel ¹, Erik Hjerde ², Lieven Sterck ^{3,4},
Salvadors Capella-Gutierrez ^{5,6}, Cederic Notredame^{7,8},
Olga Vinnere Pettersson⁹, Joelle Amselem ¹⁰, Laurent Bouri ¹,
Stephanie Bocs ¹¹⁻¹³, Christophe Klopp ¹⁴, Jean-Francois Gibrat ^{1,15},
Anna Vlasova ⁸, Brane L. Leskosek¹⁶, Lucile Soler¹⁷, Mahesh Binzer-Panchal ¹⁷,
Henrik Lantz ¹⁷

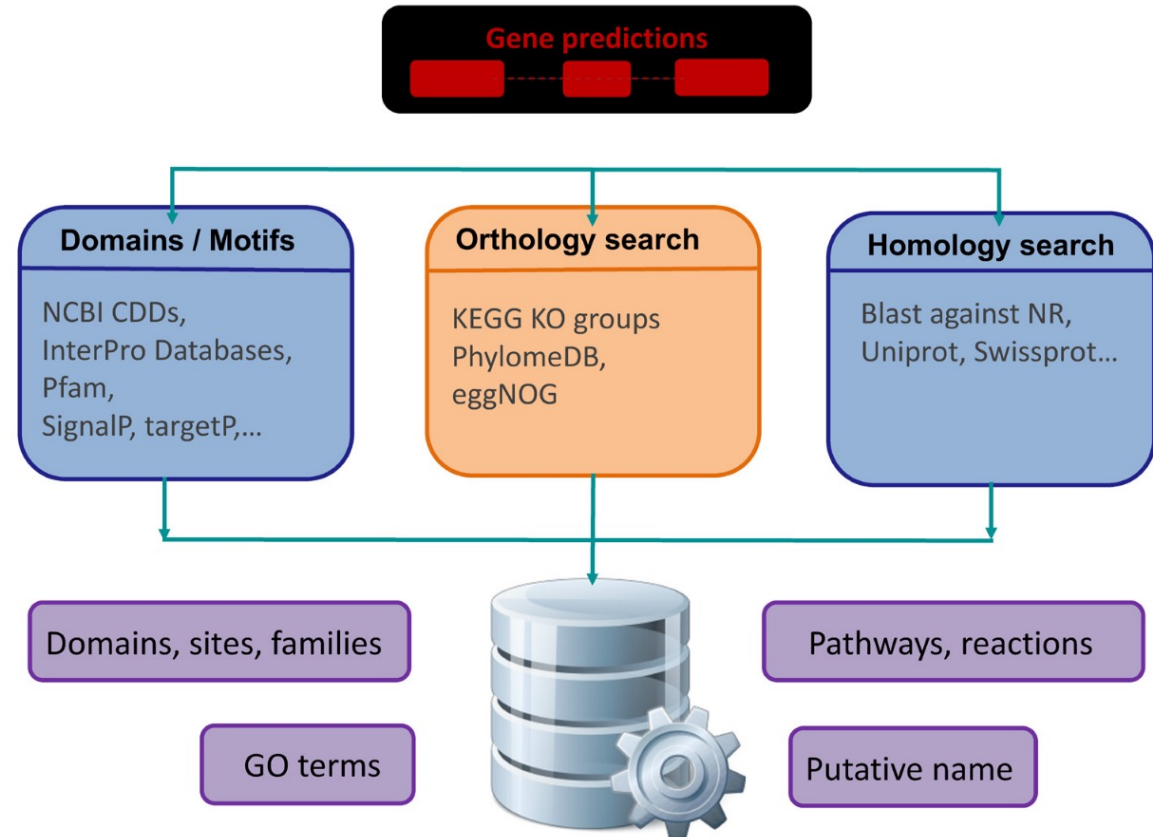
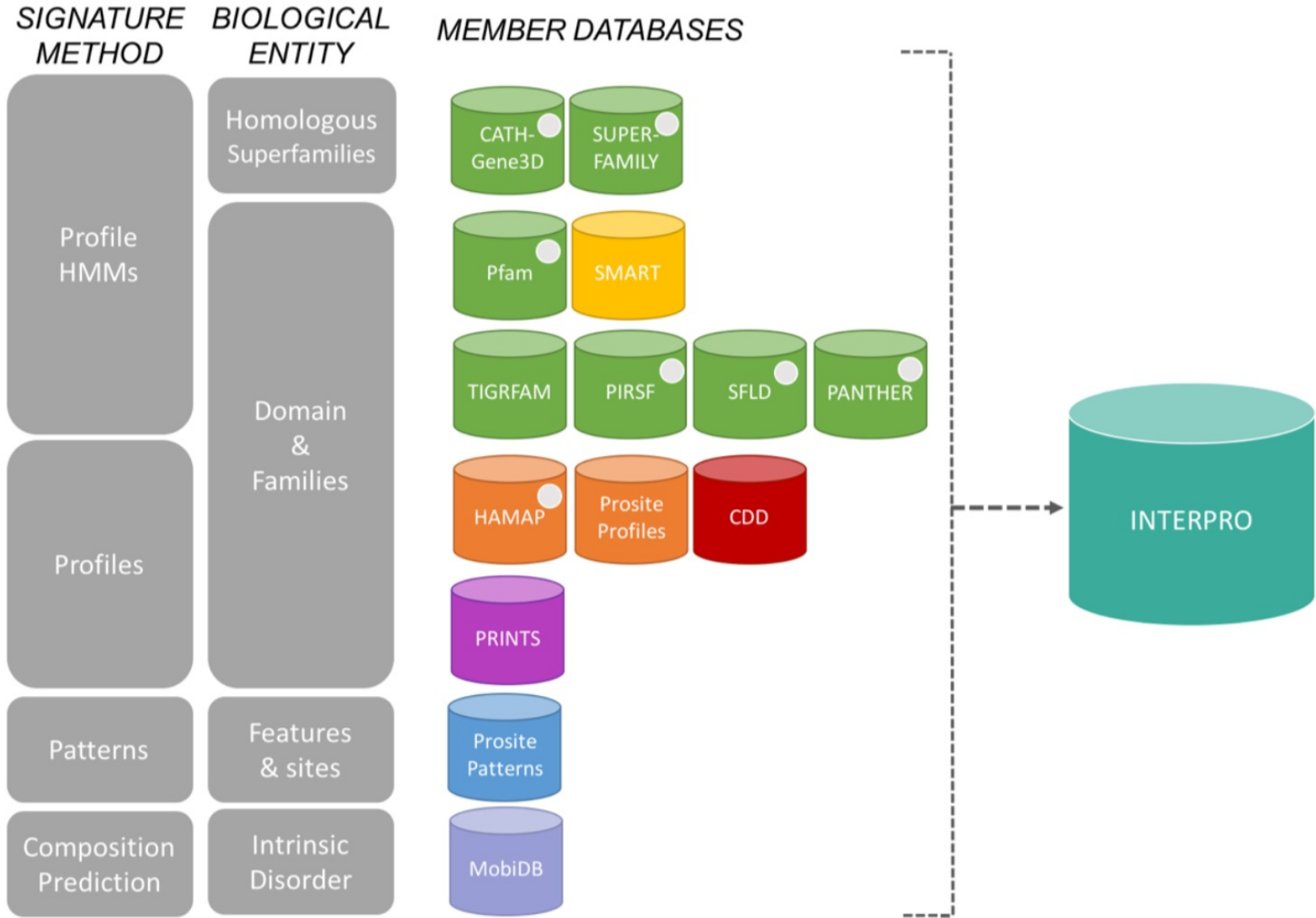


Figure 4. Functional Annotation Pipelines. This schema is showing a typical functional annotation pipeline, in which functional roles are assigned to coding sequences (CDSs) inferred in the gene prediction process. The process implements three parallel routes for the definition of functions. The first refers to proteins domains and motifs, the second for orthology search and finally the third is applied to homology search. At the end, the output from the three different sources is put together for more valuable predictions.

Domains / Motifs

NCBI CDDs,
InterPro Databases,
Pfam,
SignalP, targetP,...

Running InterProScan



InterProScan is the software package that allows sequences to be scanned against InterPro's member database signatures.

Domains / Motifs

NCBI CDDs,
InterPro Databases,
Pfam,
SignalP, targetP,...

Running InterProScan

The screenshot displays the Galaxy France web interface. At the top, the navigation bar includes the Galaxy France logo, a home icon, and menu items for Workflow, Visualize, Données partagées, Aide, Utilisateur, and a notification bell. A status indicator shows 'Using 4%'. A maintenance notice is visible below the navigation bar. The left sidebar contains a 'Tools' section with a search bar containing 'interproscan', an 'Upload Data' button, and a 'Show Sections' button. Below this is the 'InterProScan functional annotation' section, which includes a 'WORKFLOWS' tab and a list of workflows, currently showing 'All workflows'. The main content area features a large 'Welcome to usegalaxy.fr' message with a bar chart graphic and a 'Term Of Use' link. A light blue box contains a notice about the 2022 release. The right sidebar shows a 'History' section with a search bar and a list of jobs, including 'MucorProtSet' and '1983: Sort on data 1902'. The 'MucorProtSet' job is expanded, showing '2 lines' of data in a tabular format, with a table containing two columns and two rows of data.

Galaxy France

Workflow Visualize Données partagées Aide Utilisateur Using 4%

! UseGalaxy.fr will be undergoing maintenance from October 6th to 7th. Running jobs will be stopped. Thank you for your understanding.

Tools

interproscan

Upload Data

Show Sections

InterProScan functional annotation

WORKFLOWS

All workflows

Welcome to usegalaxy.fr

By using this Galaxy instance, we assume that you have read and accept the [Term Of Use](#)

For any questions or support: community.cluster.france-bioinformatique.fr/c/galaxy

- 29/06/2022: usegalaxy.fr is now running the **release 22.01** of Galaxy. Please check the [22.01 user release notes](#).

History

Rechercher des données

MucorProtSet

32 shown, 73 deleted, 1881 hidden

3.99 GB

1983: Sort on data 1902

1980: Count on data 1978

2 lines

format: **tabular**, génome de référence: ?

Count of unique values in c2

1	2
10937	EGG
11346	IPS

Domains / Motifs

NCBI CDDs,
InterPro Databases,
Pfam,
SignalP, targetP,...

Running InterProScan

Galaxy France Workflow Visualize Données partagées Aide Utilisateur

! UseGalaxy.fr will be undergoing maintenance from October 6th to 7th. Running jobs will be stopped. Thank you for your understanding.

Tools ☆ ☰

interproscan ✕

Upload Data

Show Sections

InterProScan functional annotation

WORKFLOWS

All workflows

InterProScan functional annotation (Galaxy Version 5.55-88.0+galaxy3) ☆ ☰

Protein FASTA File

📄 📄 📁 3: Galaxy13-[Funannotate_predict_annotation_on_data_4,_data_9,_and_data_6_... ⬇ 📁

(--input)

Type of the input sequences

Protein ⬇

(--seqtype)




InterProScan database

InterProScan 5.55-88.0 ⬇



Domains / Motifs


NCBI CDDs,
InterPro Databases,
Pfam,
SignalP, targetP,...


Running InterProScan


Galaxy France [Workflow](#) [Visualize](#) [Données partagées](#) [Aide](#) [Utilisateur](#)   

! UseGalaxy.fr will be undergoing maintenance from October 6th to 7th. Running jobs will be stopped. Thank you for your understanding.

Tools  

interproscan 

 **Upload Data**

 **Show Sections**

InterProScan functional annotation

WORKFLOWS

All workflows

Applications to run




Select/Unselect all

- × TIGRFAM: protein families based on hidden Markov models (HMMs)
- × SFLD: a database of protein families based on hidden Markov models (HMMs)
- × SUPERFAMILY: database of structural and functional annotation for all proteins and genomes
- × PANTHER: Protein ANALysis THrough Evolutionary Relationships
- × Gene3d: Structural assignment for whole genes and genomes using the CATH domain structure database
- × HAMAP: High-quality Automated Annotation of Microbial Proteomes
- × PROSITE Profiles: protein domains, families and functional sites as well as associated profiles to identify them
- × Coils: Prediction of Coiled Coil Regions in Proteins
- × SMART: identification and analysis of domain architectures based on Hidden Markov Models or HMMs
- × SMART: protein domains and families based on well-annotated multiple sequence alignment models
- × PRINTS: group of conserved motifs (fingerprints) used to characterise a protein family
- × PIRSR: protein families based on hidden Markov models (HMMs) and Site Rules
- × PROSITE Pattern: protein domains, families and functional sites as well as associated patterns to identify them
- × AntiFam: a resource of profile-HMMs designed to identify spurious protein predictions.
- × Pfam: protein families, each represented by multiple sequence alignments and hidden Markov models
- × MobiDBLite: Prediction of intrinsically disordered regions in proteins
- × PIRSF: non-overlapping clustering of UniProtKB sequences into a hierarchical order (evolutionary relationships)



Domains / Motifs


NCBI CDDs,
InterPro Databases,
Pfam,
SignalP, targetP,...


Running InterProScan


Galaxy France [Workflow](#) [Visualize](#) [Données partagées](#) [Aide](#) [Utilisateur](#)   

! UseGalaxy.fr will be undergoing maintenance from October 6th to 7th. Running jobs will be stopped. Thank you for your understanding.

Tools  

interproscan 

 **Upload Data**

 **Show Sections**


InterProScan functional annotation

WORKFLOWS

All workflows

Select your program

Use applications with restricted license, only for non-commercial use?

No 

The corresponding tools must be installed manually by the administrator of this Galaxy instance

Include pathway information

Yes

Option that provides mappings from matches to pathway information, which is based on the matched manually curated InterPro entries. (--pathways)

Include Gene Ontology (GO) mappings

Yes

Look up of corresponding Gene Ontology annotation. Implies -iplookup option. (--goterms)

Provide additional mappings

No

Provide mappings from matched member database signatures to the InterPro entries that they are integrated into (--iplookup)

Output format

Select/Unselect all


Tab-separated values format (TSV) GFF3 XML JSON

Please select a output format (JSON output can be visualised on <https://www.ebi.ac.uk/interpro/result/InterProScan/>).

Email notification

No

Send an email notification when the job completes.

 **Execute**

Domains / Motifs

NCBI CDDs,
InterPro Databases,
Pfam,
SignalP, targetP,...

Running InterProScan

Galaxy France

Workflow Visualize Données partagées Aide Utilisateur Using 4%

! UseGalaxy.fr will be undergoing maintenance from October 6th to 7th. Running jobs will be stopped. Thank you for your understanding.

Tools

interproscan

Upload Data

Show Sections

InterProScan functional annotation

WORKFLOWS

All workflows

Executed **InterProScan** and successfully added 1 job to the queue.

The tool uses this input:

- 3: Galaxy13-[Funannotate_predict_annotation_on_data_4,_data_9,_and_data_6__protein_sequences].fasta

It produces 4 outputs:

- 1984: InterProScan on data 3 (tsv)
- 1985: InterProScan on data 3 (xml)
- 1986: InterProScan on data 3 (gff3)
- 1987: InterProScan on data 3 (json)

You can check the status of queued jobs and view the resulting data by refreshing the History panel. When the job has been run the status will change from 'running' to 'finished' if completed successfully or 'error' if problems were encountered.

History

Rechercher des données

MucorProtSet

36 shown, 73 deleted, 1881 hidden

3.99 GB

- 1987: InterProScan on data 3 (json)
- 1986: InterProScan on data 3 (gff3)
- 1985: InterProScan on data 3 (xml)
- 1984: InterProScan on data 3 (tsv)
- 1983: Sort on data 1902

Functional annotation pipelines



OPINION ARTICLE

Ten steps to get started in Genome Assembly and Annotation

[version 1; peer review: 2 approved]

Victoria Dominguez Del Angel ¹, Erik Hjerde ², Lieven Sterck ^{3,4},
Salvadors Capella-Gutierrez ^{5,6}, Cederic Notredame^{7,8},
Olga Vinnere Pettersson⁹, Joelle Amselem ¹⁰, Laurent Bouri ¹,
Stephanie Bocs ¹¹⁻¹³, Christophe Klopp ¹⁴, Jean-Francois Gibrat ^{1,15},
Anna Vlasova ⁸, Brane L. Leskosek¹⁶, Lucile Soler¹⁷, Mahesh Binzer-Panchal ¹⁷,
Henrik Lantz ¹⁷

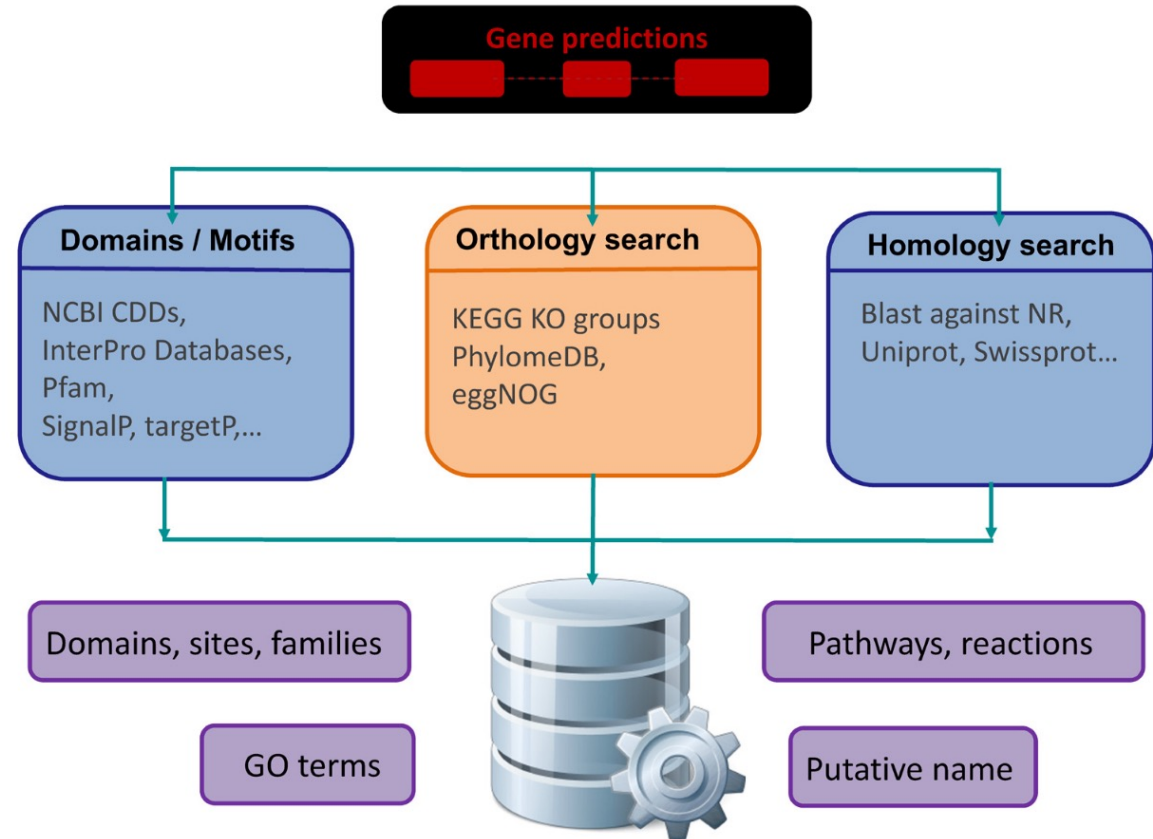


Figure 4. Functional Annotation Pipelines. This schema is showing a typical functional annotation pipeline, in which functional roles are assigned to coding sequences (CDSs) inferred in the gene prediction process. The process implements three parallel routes for the definition of functions. The first refers to proteins domains and motifs, the second for orthology search and finally the third is applied to homology search. At the end, the output from the three different sources is put together for more valuable predictions.

Orthology search

KEGG KO groups
PhylomeDB,
eggNOG

Running EggNOG-mapper

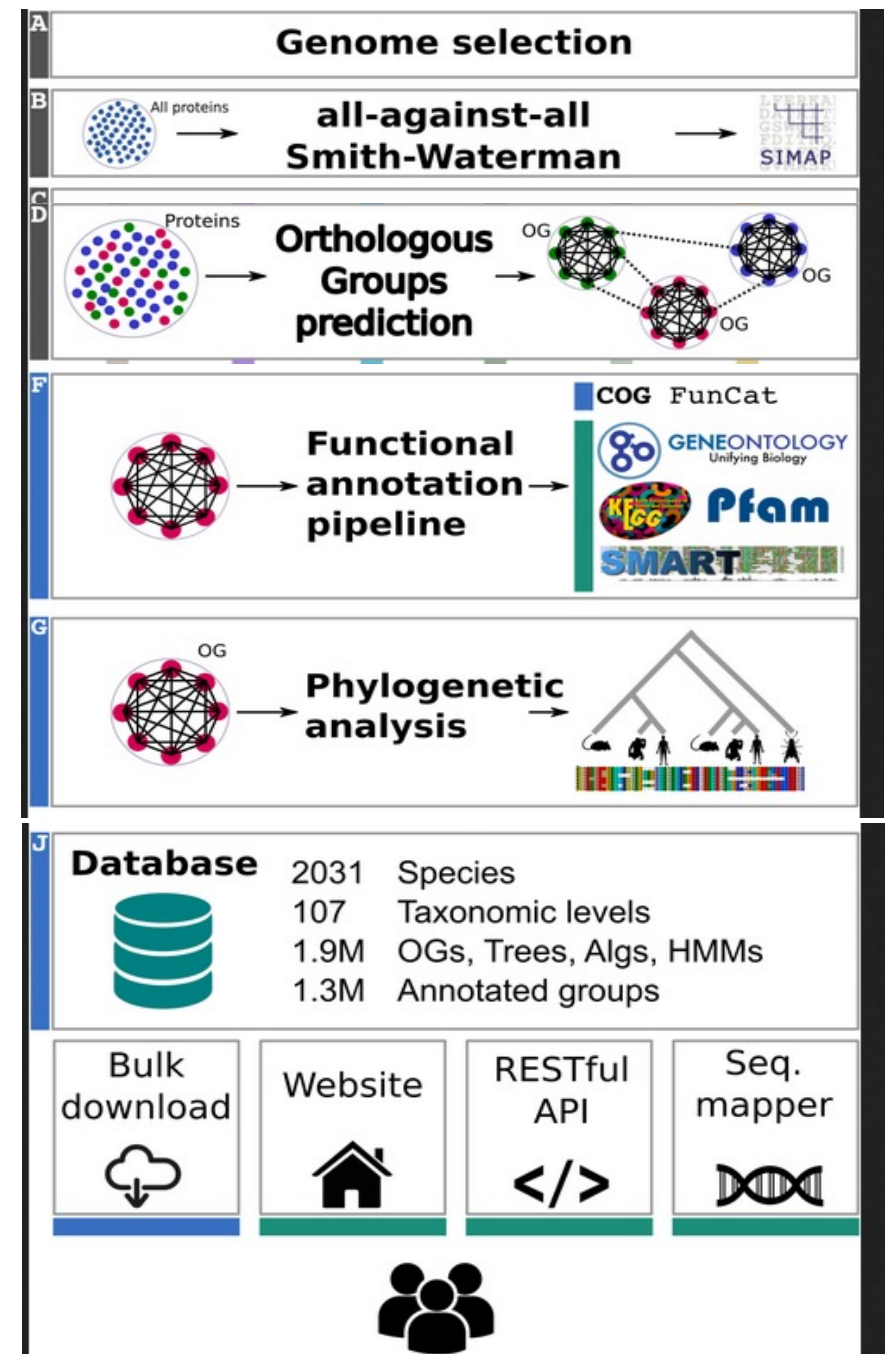
EggNOG v5.0

A database of orthology relationships, functional annotation, and gene evolutionary histories.

Organisms	Viruses	Orthologous Groups	Tree & Algs
5,090	2,502	4.4M	4.4M

EggNOG-mapper is a tool for fast functional annotation of novel sequences. It uses precomputed Orthologous Groups (OGs) and phylogenies from the EggNOG database (<http://eggnog5.embl.de>) to transfer functional information from fine-grained orthologs only.

evolutionary genealogy of genes: Non-supervised Orthologous Groups



Orthology search

KEGG KO groups
PhylomeDB,
eggNOG

Running EggNOG-mapper

Galaxy France Workflow Visualize Données partagées Aide Utilisateur

! UseGalaxy.fr will be undergoing maintenance from October 6th to 7th. Running jobs will be stopped. Thank you for your understanding.

Tools ☆ ☰

eggnog ✕

Upload Data

Show Sections

eggNOG Mapper functional sequence annotation by orthology

Funannotate functional annotation

WORKFLOWS

All workflows

eggNOG Mapper functional sequence annotation by orthology (Galaxy Version 2.1.8+galaxy3) ☆ ⚙

Version of eggNOG Database

5.0.2

Method to search seed orthologs

Diamond

(-m)

Fasta sequences to annotate

3: Galaxy13-[Funannotate_predict_annotation_on_data_4,_data_9,_and_data_6__protein_sequences].f... ⬆️ 📁

(-i)

Type of sequences

proteins

(--itype)

Scoring matrix and gap costs

BLOSUM62

(--matrix)

Gap Costs

Existence: 11 Extension: 1

Diamond's sensitivity mode

sensitive

Orthology search

KEGG KO groups
PhylomeDB,
eggNOG

Galaxy France

Workflow Visualize Données partagées Aide Utilisateur

! UseGalaxy.fr will be undergoing maintenance from October 6th to 7th. Running jobs will be stopped. Thank you for your understanding.

Tools

eggnog

Upload Data

Show Sections

eggnOG Mapper functional sequence annotation by orthology

Funannotate functional annotation

WORKFLOWS

All workflows

Min E-value threshold

0,001

Min E-value expected when searching for seed eggNOG ortholog. Applies to phmmer/diamond searches. Queries not having a significant seed orthologs (E-value less than threshold) will not be annotated. (`--seed_ortholog_evalue`)

Minimum bit score threshold

Min bit score expected when searching for seed eggNOG ortholog. Queries not having a significant seed orthologs will not be annotated. (`--seed_ortholog_score`)

Set taxonomic scope

NCBI taxonomy id (`--tax_scope`)

target orthologs for functional transfer

all

(`--target_orthologs`)

Select the set of GO terms that should be used for annotation

non-electronic = Use only non-electronically curated terms

(`--go_evidence`)

Output Options

Email notification

No

Send an email notification when the job completes.

Execute

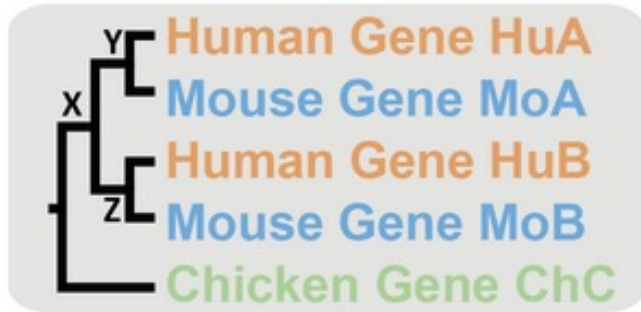
To be continued ...

How to assign a biological function to a gene/protein ?

- Experimental characterization
- Automated prediction:
 - 1/ Search for similar genes/proteins in other organisms
 - 2/ Transfer the annotation to the query sequence

The assumption of function transfer is that the function is retained in proteins that have similar sequences and have evolved from a single ancestor.

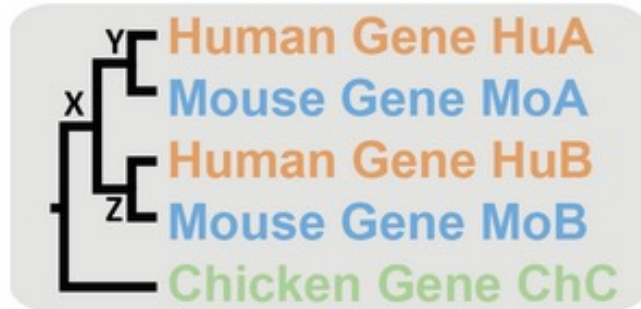
How genes evolve ?



**Group of genes descended
from single gene in LCA
of group of species**

How genes evolve ?

A. Orthogroup



Group of genes descended
from single gene in LCA
of group of species

How genes evolve ?

A. Orthogroup



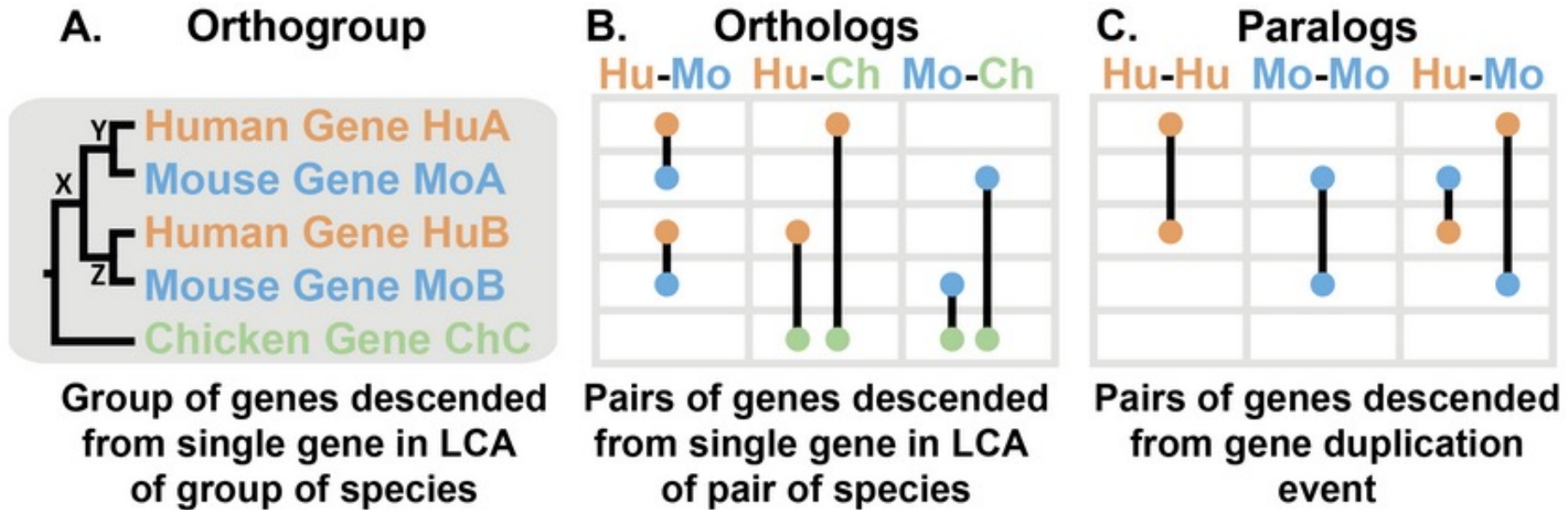
Group of genes descended from single gene in LCA of group of species

B. Orthologs



Pairs of genes descended from single gene in LCA of pair of species

How genes evolve ?



LCA: Last Common Ancestor

How genes evolve ?

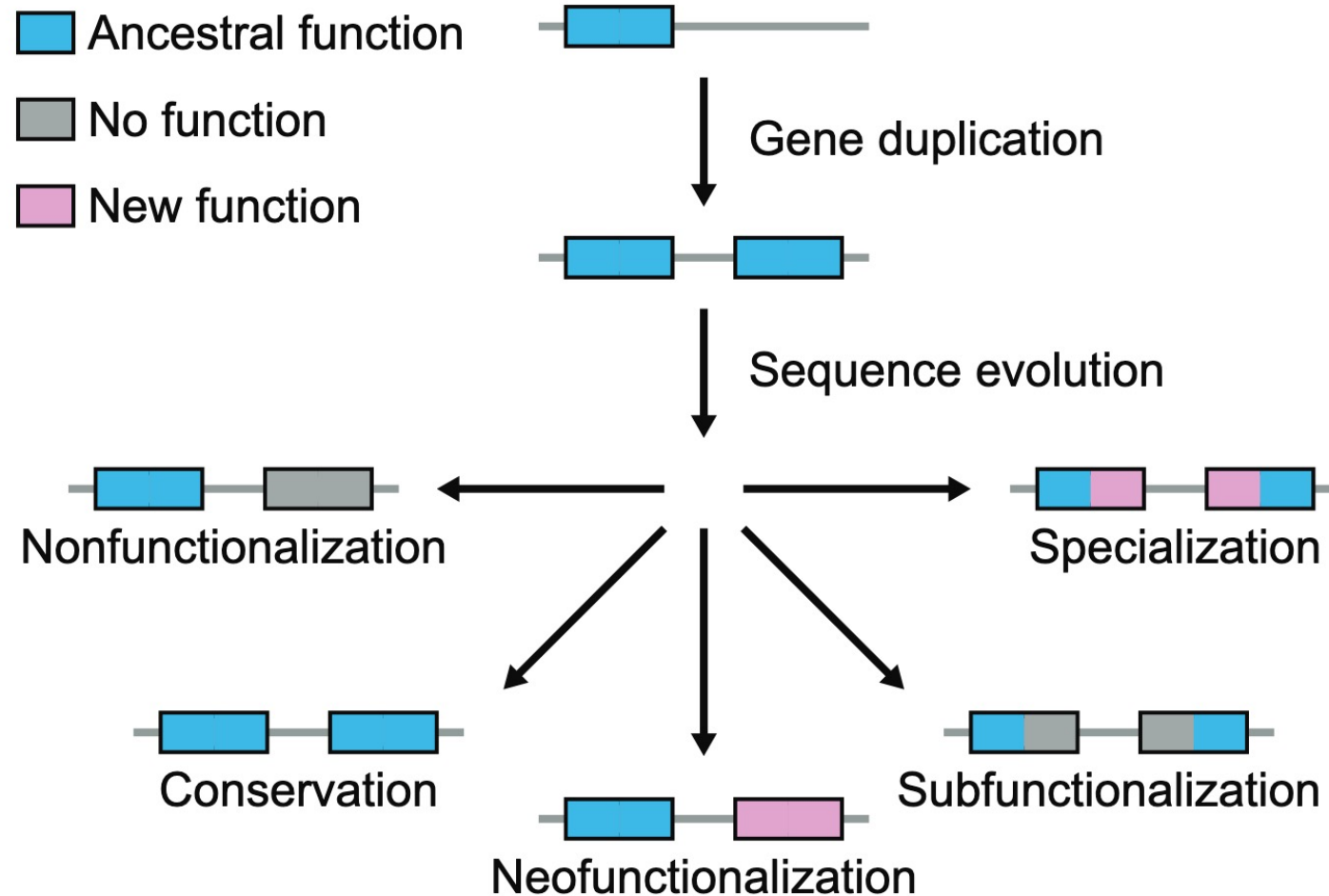
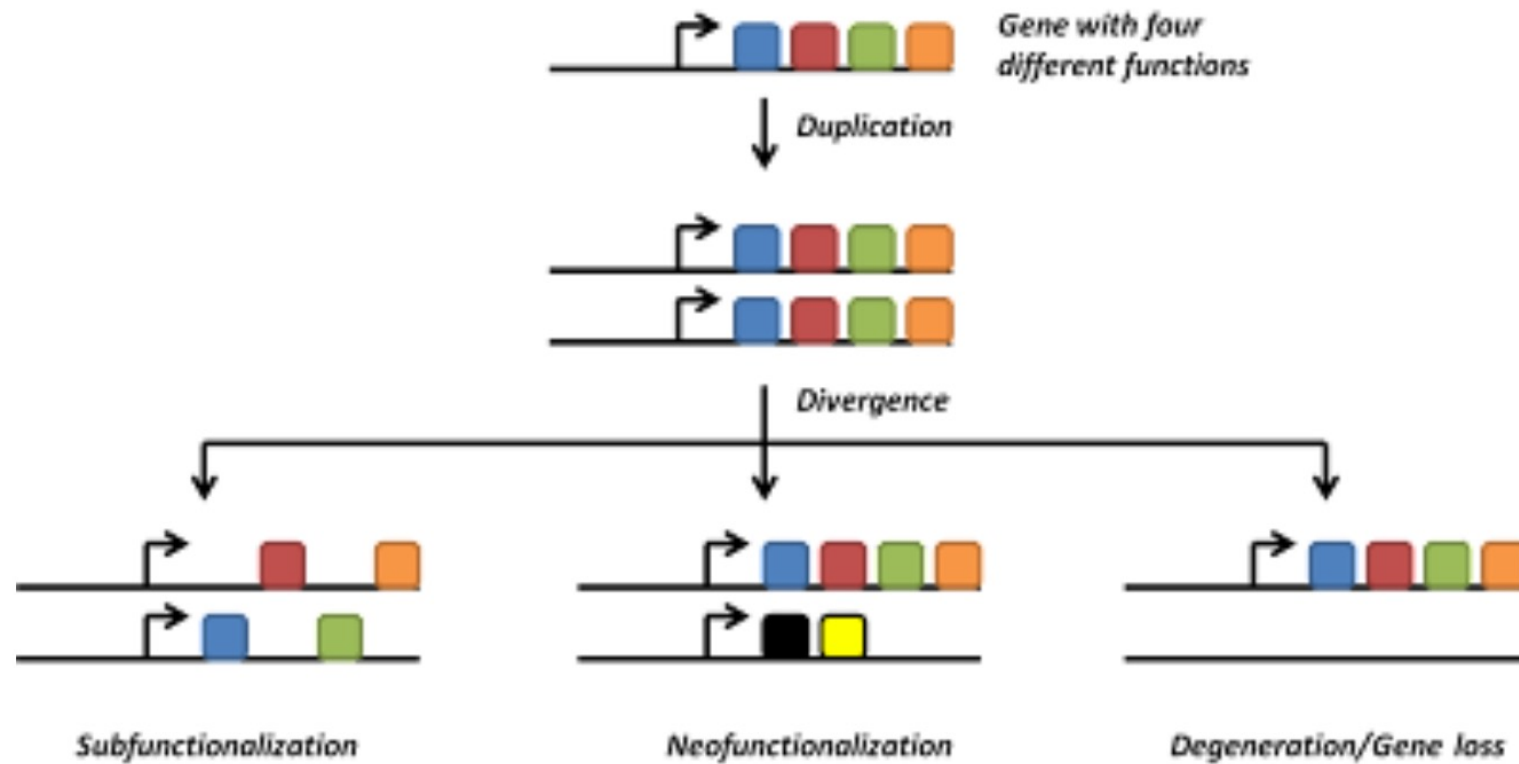


Figure 1: Hypothesized evolutionary trajectories of duplicate genes. Gene duplication results in two copies of an ancestral gene. Evolution may result in the loss of one functional copy by nonfunctionalization, or in the retention of two functional copies by either conservation, neofunctionalization, subfunctionalization, or specialization.

How genes evolve ?

Evolutionary fate of duplicate genes



Search for sequence similarity BLAST

NIH National Library of Medicine
National Center for Biotechnology Information

Log in

BLAST® Home Recent Results Saved Strategies Help

Basic Local Alignment Search Tool

BLAST finds regions of similarity between biological sequences. The program compares nucleotide or protein sequences to sequence databases and calculates the statistical significance. [Learn more](#)

NEWS
BLAST+ 2.13.0 is here!
Starting with this release, we are including the blastn_vdb and tblastn_vdb executables in the BLAST+ distribution.
Thu, 17 Mar 2022 12:00:00 EST [More BLAST news...](#)

Web BLAST

Nucleotide BLAST
nucleotide ▶ nucleotide

blastx
translated nucleotide ▶ protein

tblastn
protein ▶ translated nucleotide

Protein BLAST
protein ▶ protein

Which database ? (quality redundancy ...)
Which cut-off ? (E-value)

Local alignment searches suffer from some important caveats

Local alignment searches suffer from some important caveats:

1. Excessive transfer of annotations.



2. Annotation errors in the source database.

3. Threshold relativity.

4. Low sensitivity/specificity.

5. Paralogs versus orthologs.

AUTOMATED FUNCTION PREDICTION

Functional annotation prediction: All for one and one for all

ORI SASSON,^{1,3} NOAM KAPLAN,^{2,3} AND MICHAL LINIAL²

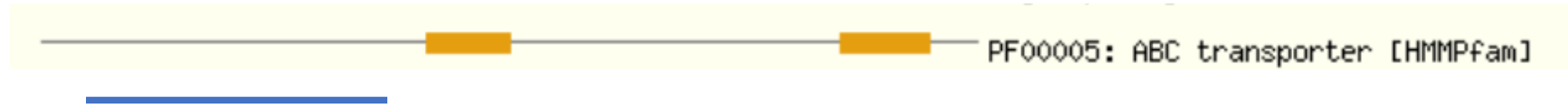
¹School of Computer Science and Engineering, and ²Department of Biological Chemistry, The Hebrew University of Jerusalem, Jerusalem 91904, Israel

(RECEIVED February 23, 2006; FINAL REVISION February 23, 2006; ACCEPTED February 23, 2006)

Local alignment searches suffer from some important caveats

Local alignment searches (BLAST) suffer from some important caveats:

1. Excessive transfer of annotations.



2. Annotation errors in the source database.

3. Threshold relativity.

4. Low sensitivity/specificity.

5. Paralogs versus orthologs.

6. Retrieve essentially poorly formatted text

AUTOMATED FUNCTION PREDICTION

Functional annotation prediction: All for one and one for all

ORI SASSON,^{1,3} NOAM KAPLAN,^{2,3} AND MICHAL LINIAL²

¹School of Computer Science and Engineering, and ²Department of Biological Chemistry, The Hebrew University of Jerusalem, Jerusalem 91904, Israel

(RECEIVED February 23, 2006; FINAL REVISION February 23, 2006; ACCEPTED February 23, 2006)

Functional annotation pipelines



OPINION ARTICLE

Ten steps to get started in Genome Assembly and Annotation

[version 1; peer review: 2 approved]

Victoria Dominguez Del Angel ¹, Erik Hjerde ², Lieven Sterck ^{3,4},
Salvadors Capella-Gutierrez ^{5,6}, Cederic Notredame^{7,8},
Olga Vinnere Pettersson⁹, Joelle Amselem ¹⁰, Laurent Bouri ¹,
Stephanie Bocs ¹¹⁻¹³, Christophe Klopp ¹⁴, Jean-Francois Gibrat ^{1,15},
Anna Vlasova ⁸, Brane L. Leskosek¹⁶, Lucile Soler¹⁷, Mahesh Binzer-Panchal ¹⁷,
Henrik Lantz ¹⁷

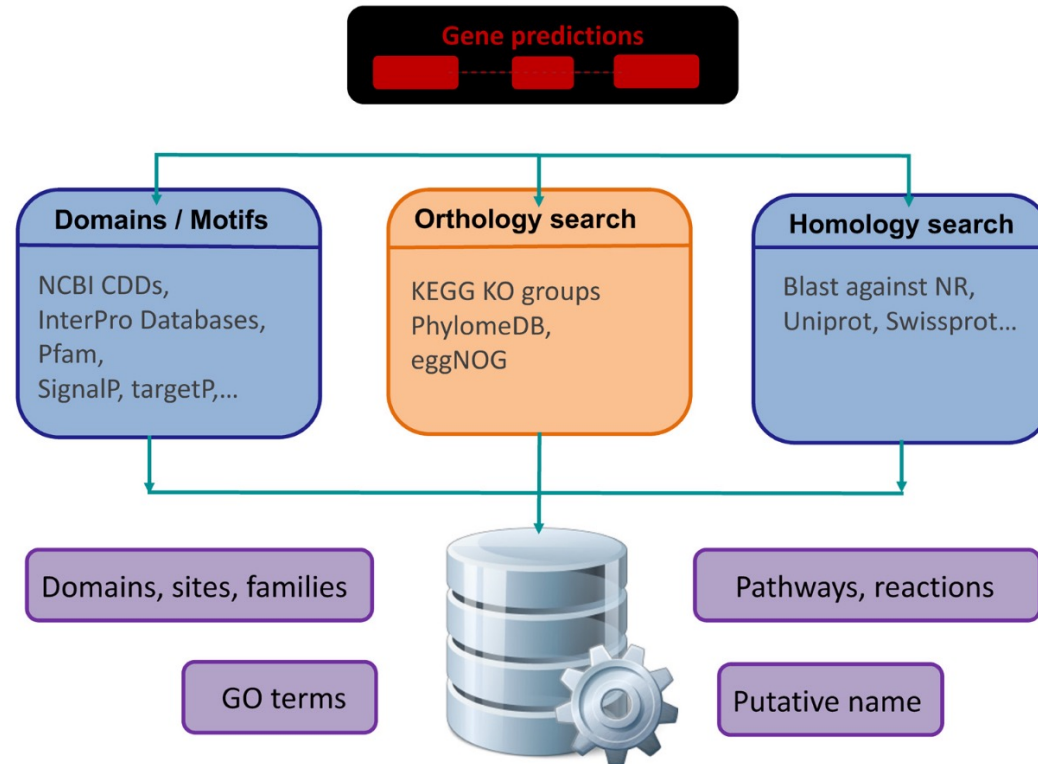


Figure 4. Functional Annotation Pipelines. This schema is showing a typical functional annotation pipeline, in which functional roles are assigned to coding sequences (CDSs) inferred in the gene prediction process. The process implements three parallel routes for the definition of functions. The first refers to proteins domains and motifs, the second for orthology search and finally the third is applied to homology search. At the end, the output from the three different sources is put together for more valuable predictions.

Proteins classification

Proteins can be classified into groups according to sequence or structural similarity.

These groups often contain well characterised proteins whose function is known.

Thus, when a novel protein is identified, its functional properties can be proposed based on the group to which it is predicted to belong.

Proteins can be classified into groups based on

- the **FAMILIES** to which they belong
- the **DOMAINS** they contain
- the **SEQUENCE FEATURES** they possess

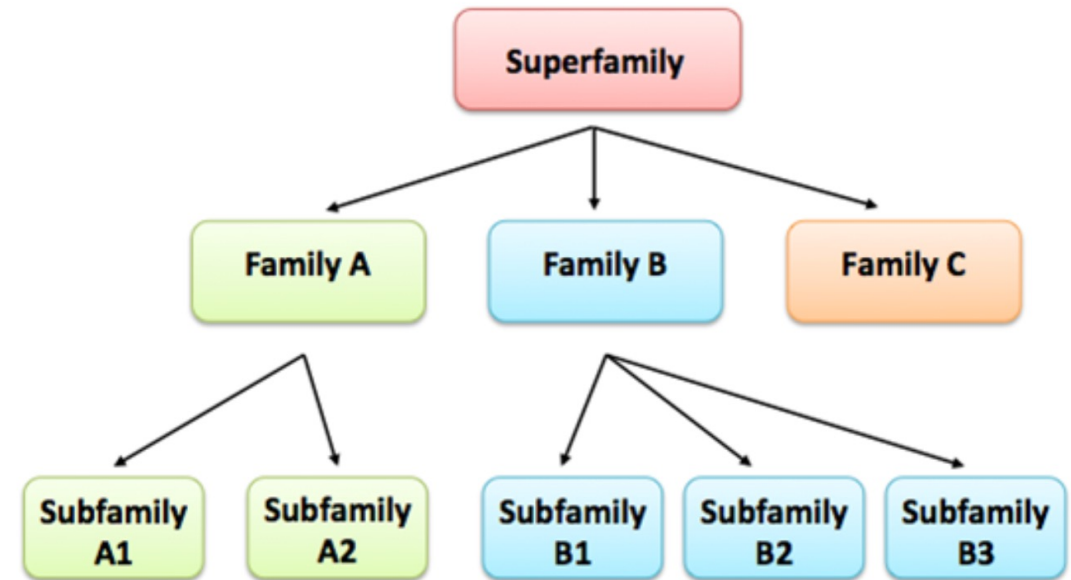


Figure 2 A hypothetical protein family hierarchy showing the relationships between superfamily, family and subfamily members. Directional arrows indicate that one group is a subgroup of another.

Protein families

Families

A protein family is a group of proteins that share a common evolutionary origin, reflected by their related functions and similarities in sequence or structure.

Protein families are often arranged into hierarchies, with proteins that share a common ancestor subdivided into smaller, more closely related groups.

The terms **superfamily** (describing a large group of distantly related proteins) and **subfamily** (describing a small group of closely related proteins) are sometimes used in this context.

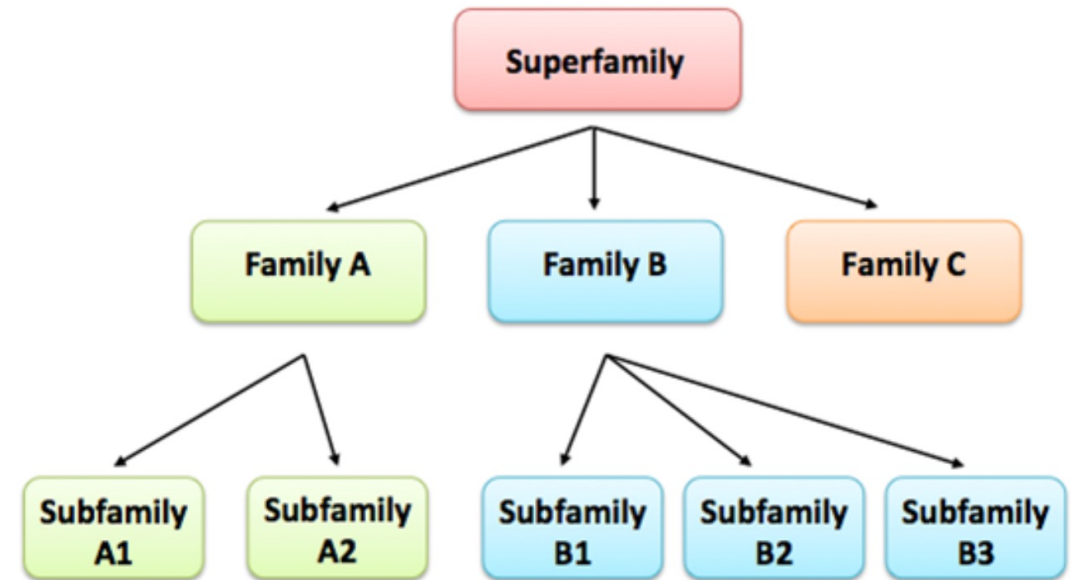


Figure 2 A hypothetical protein family hierarchy showing the relationships between superfamily, family and subfamily members. Directional arrows indicate that one group is a subgroup of another.

Protein domains

Domains

Domains are distinct functional and/or structural units in a protein.

Usually they are responsible for a particular function or interaction, contributing to the overall role of a protein.

Domains may exist in a variety of biological contexts, where similar domains can be found in proteins with different functions

Protein family/subfamily can often be defined by an specific domain arrangement

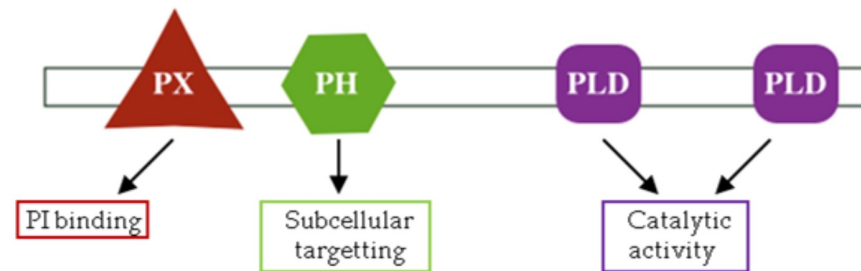
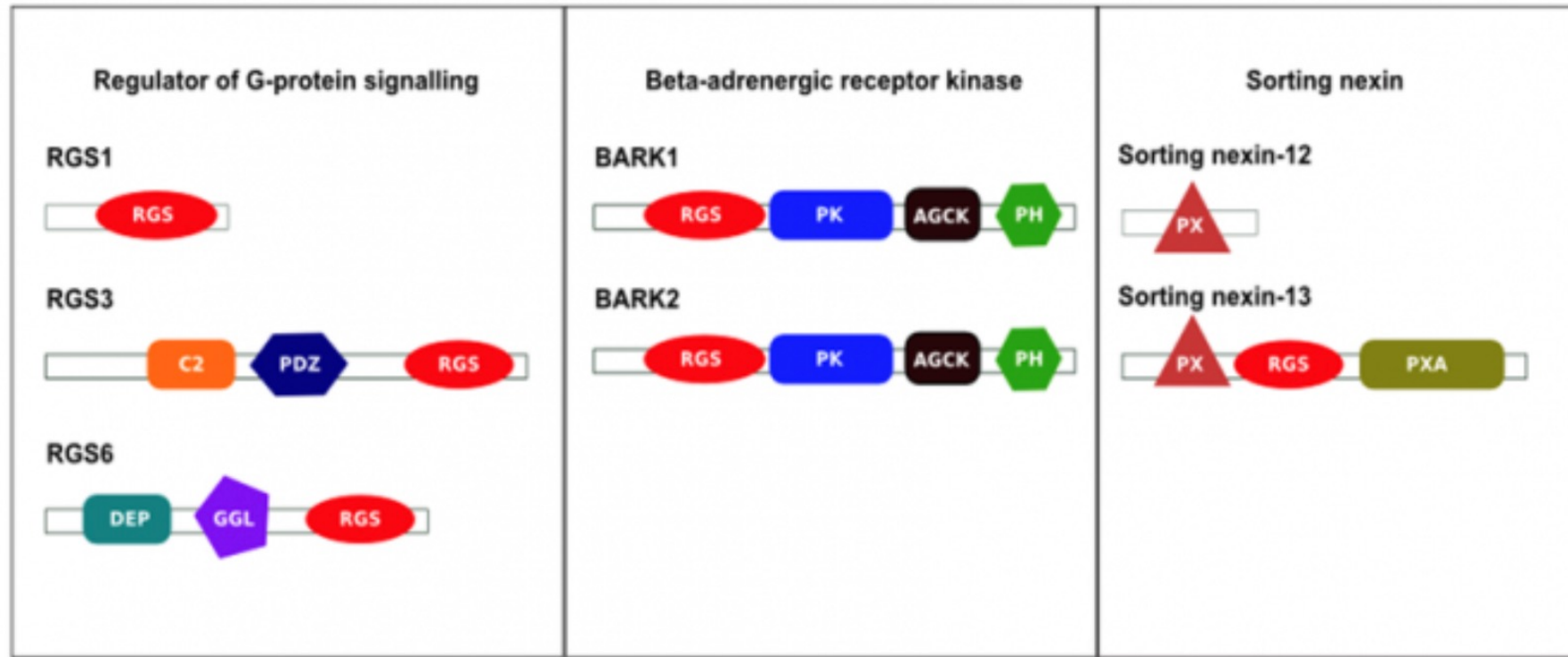


Figure 6 Domain composition of phospholipase D1, which is an enzyme that breaks down phosphatidylcholine.

Protein domains

Family- and domain-based protein classification

Family- and domain-based classifications are not always straightforward and can overlap, since proteins are sometimes assigned to families by virtue of the domain(s) they contain.

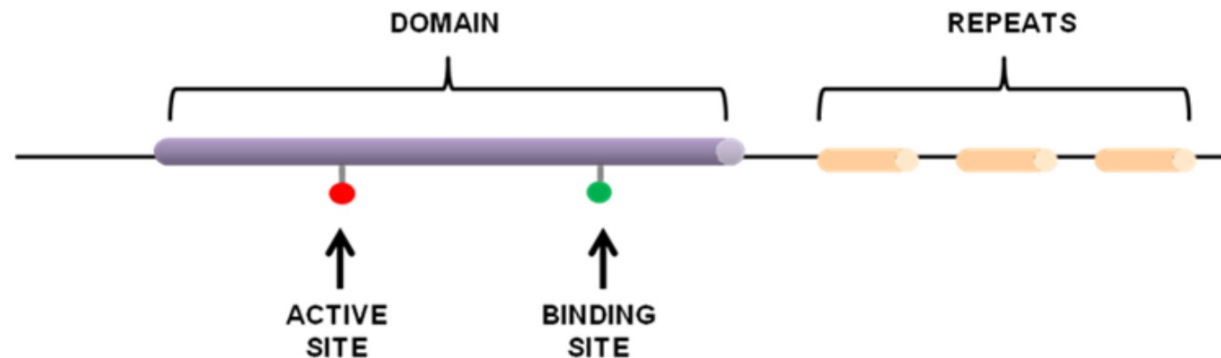


Protein sequence features

Sequence features are groups of amino acids that confer certain characteristics upon a protein

- **active sites.**
- **binding sites**
- **post-translational modification (PTM) sites.**
- **repeats.**

Sequence features differ from domains in that they are usually quite small (often only a few amino acids long), whereas domains represent entire structural or functional units of the protein



Protein domains and features have signatures

To classify proteins into families and to predict the presence of important domains or sequence features, we need predictive models known as **protein signatures**.

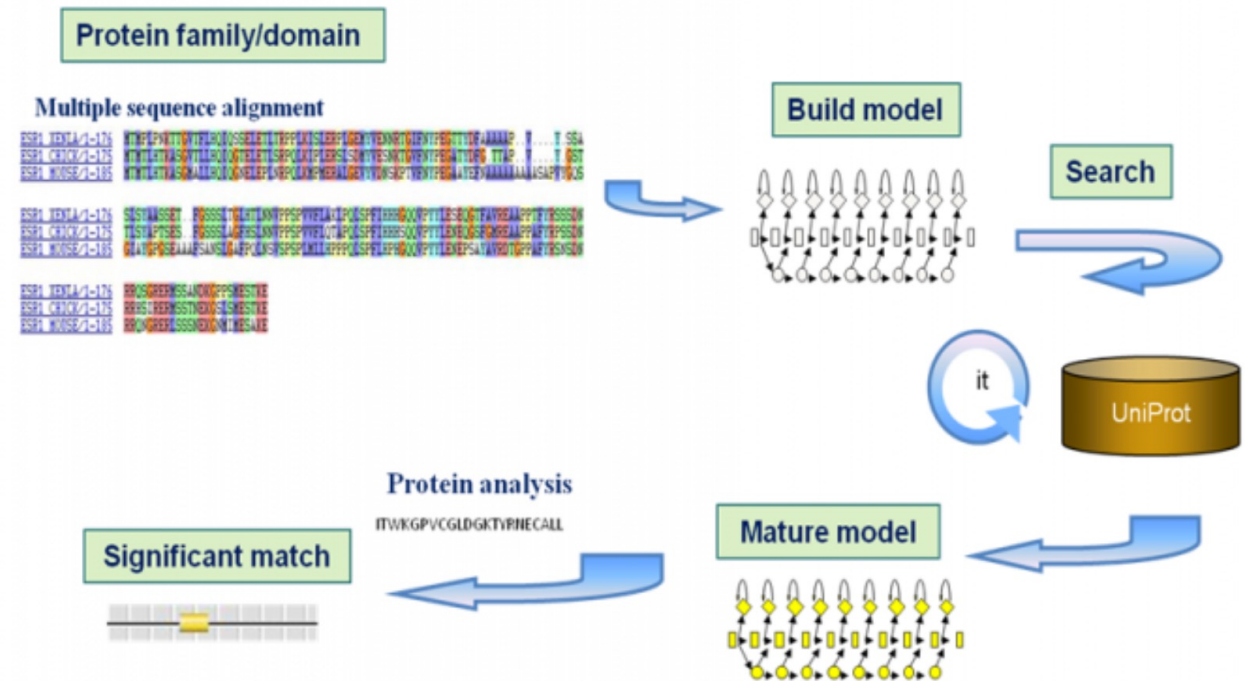
There are different types of signatures, built using different computational approaches.

A common starting point is a multiple sequence alignment of proteins sharing a set of characteristics (e.g. belonging to the same family or sharing a domain).

1/ When building the initial model, the level of amino acid conservation at different positions in the alignment is taken into account.

2/ The model is then used to search a protein database in an iterative manner, refining the model as more distantly related sequences in the database are identified.

3/ Once the model is mature, the signature is ready and can be used for protein sequence analysis.

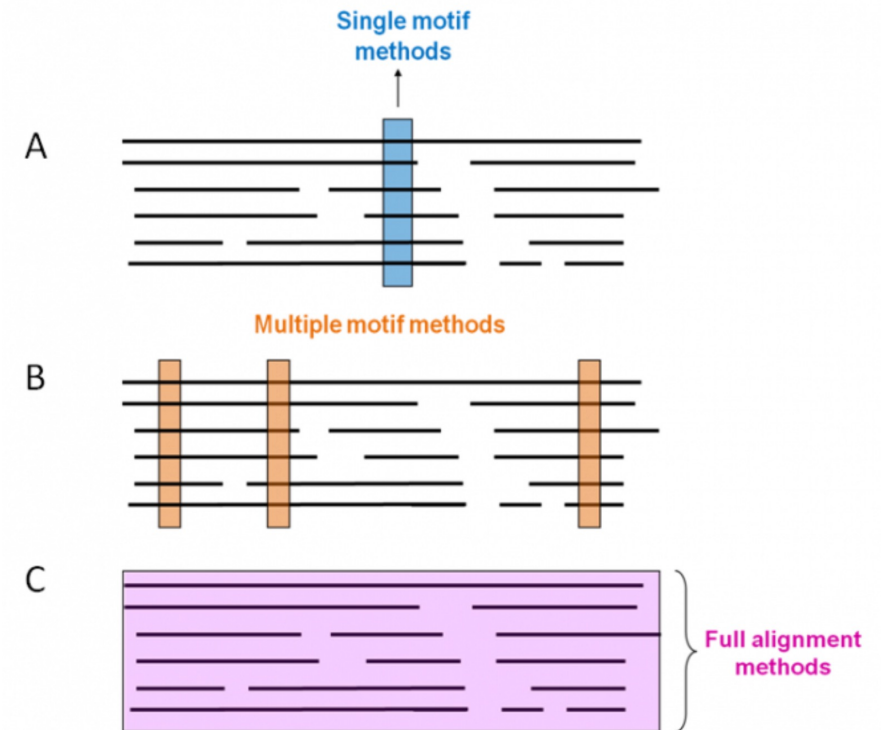


Protein signature types

Different approaches can be used to generate signatures. These include:

- patterns
- profiles
- fingerprints
- hidden Markov models (HMMs)

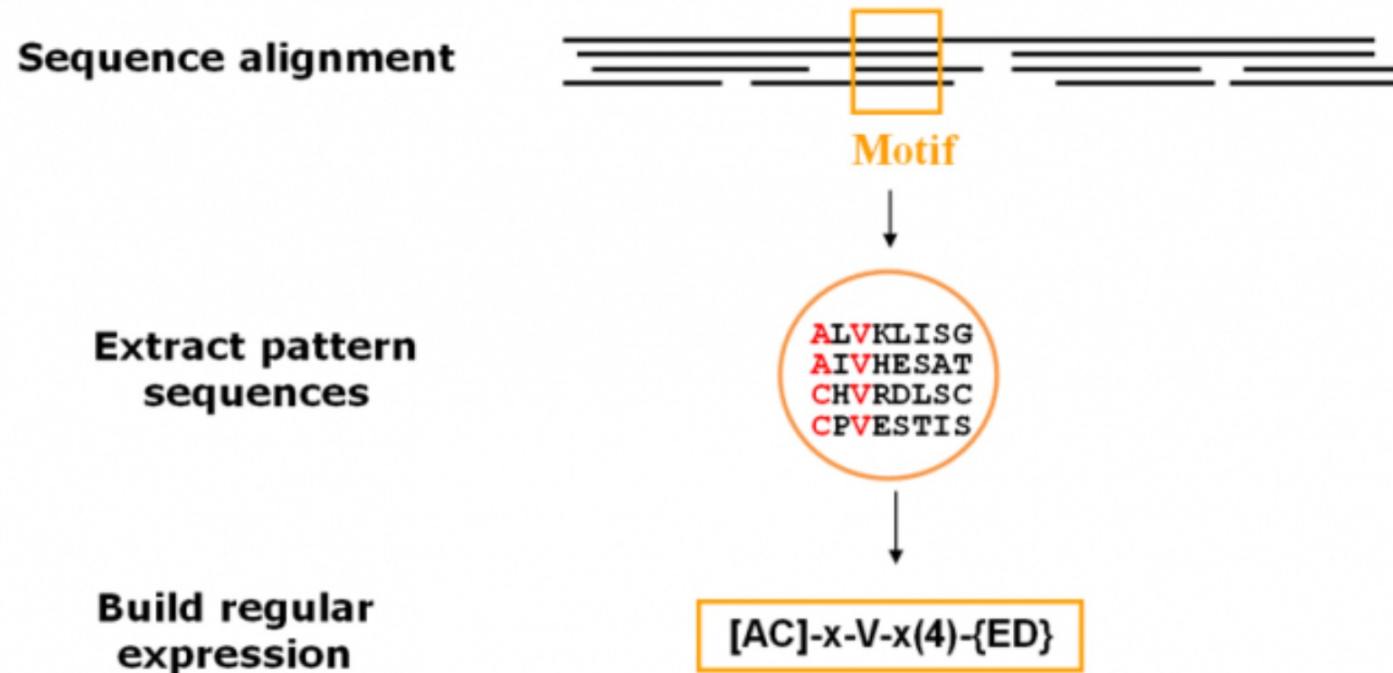
Each approach starts with a protein multiple sequence alignment, and can focus on a **single** conserved sequence region (known as a motif), **multiple** conserved motifs, or the **full** alignment of the entire protein or a particular domain .



Protein patterns

Many important sequence features, such as binding sites or the active sites of enzymes, consist of only a few amino acids that are essential for protein function.

The pattern of conservation within the sequence feature is then modelled as a **regular expression**. An example of a database that uses patterns is **PROSITE**.



Protein profiles

Profiles are used to model protein families and domains.

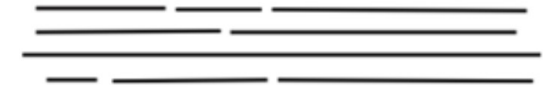
They are built by converting multiple sequence alignments into **position-specific scoring systems** (PSSMs).

Amino acids at each position in the alignment are scored according to the frequency with which they occur.

Examples of databases

- **CDD**
- **HAMAP**
- **PROSITE** (which produces profiles as well as patterns)
- **The PRODOM** database also uses a related approach, using PSI-BLAST to create its profiles.

Sequence alignment



Residue frequency at each position

Sequence 1:	F	K	L	L	S	H	C	L	L	V
Sequence 2:	F	K	A	P	G	Q	T	M	P	Q
Sequence 3:	Y	P	I	V	G	Q	E	L	L	G
Sequence 4:	F	P	V	V	K	E	A	I	L	K
Sequence 5:	F	K	V	L	A	A	V	I	A	D
Sequence 6:	L	E	F	I	S	E	C	I	I	Q
Sequence 7:	F	K	L	L	G	N	V	L	V	C



Scoring matrix

A	-18	-10	-1	-8	8	-3	3	-10	-2	-8
C	-22	-32	-18	-18	-22	-26	22	-24	-19	-7
D	-25	0	-32	-23	-7	6	-17	-34	-31	0
E	-27	15	-25	-26	-9	23	-9	-24	-23	-1
F	60	-30	12	14	-26	-29	-15	4	12	-29
G	-30	-20	-28	-32	28	-14	-23	-33	-27	-5
H	-13	-12	-25	-25	-16	14	-22	-22	-23	-10
I	3	-27	21	25	-29	-23	-8	33	19	-23
K	-24	25	-25	-27	-4	4	-15	-27	-26	0
L	14	-28	19	27	-27	-20	-9	33	26	-21
M	3	-15	10	14	-17	-10	-9	25	12	-11
N	-22	-6	-24	-27	1	8	-15	-24	-24	-4
P	-30	24	-24	-26	-14	-10	-22	-24	-26	-10
Q	-32	8	-25	-24	-9	24	-16	-17	-23	7
R	-18	9	-22	-22	-10	0	-18	-23	-22	-4
S	-22	-8	-16	-21	11	2	-1	-24	-19	-4
T	-10	-10	-6	-7	-3	-8	2	-10	-7	-11
V	0	-25	22	25	-19	-26	6	19	16	-14
W	9	-25	-18	-19	-25	-27	-34	-20	-17	-28
Y	34	-18	-1	1	-23	-12	-19	0	0	-18

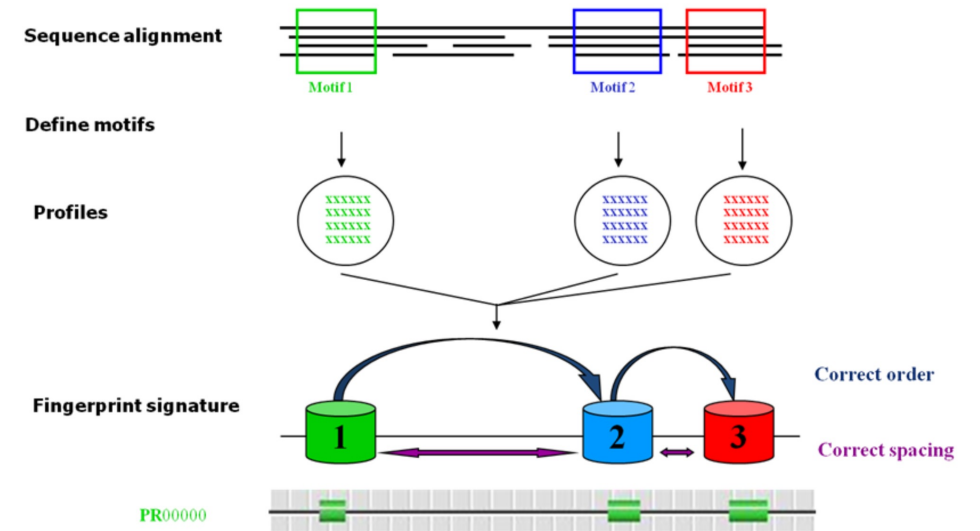
Protein fingerprints

Most protein families are characterised not by one, but by several conserved regions, which occur in a certain order.

Fingerprints are composed of multiple short conserved motifs. Each motif is converted into an individual profile (as described in the previous slide) to create a fingerprint signature.

Fingerprints can distinguish individual subfamilies

Exemple of database: PRINTS



Hidden Markov Models

Hidden Markov models (HMMs) can be used to convert multiple sequence alignments into position-specific scoring systems.

HMMs are adept at representing amino acid insertions and deletions, **meaning that they can model entire alignments, including divergent regions.**

Databases:

- Pfam
- SMART
- TIGRFAM
- PIRSF
- PANTHER
- SFLD
- Superfamily
- Gene3D

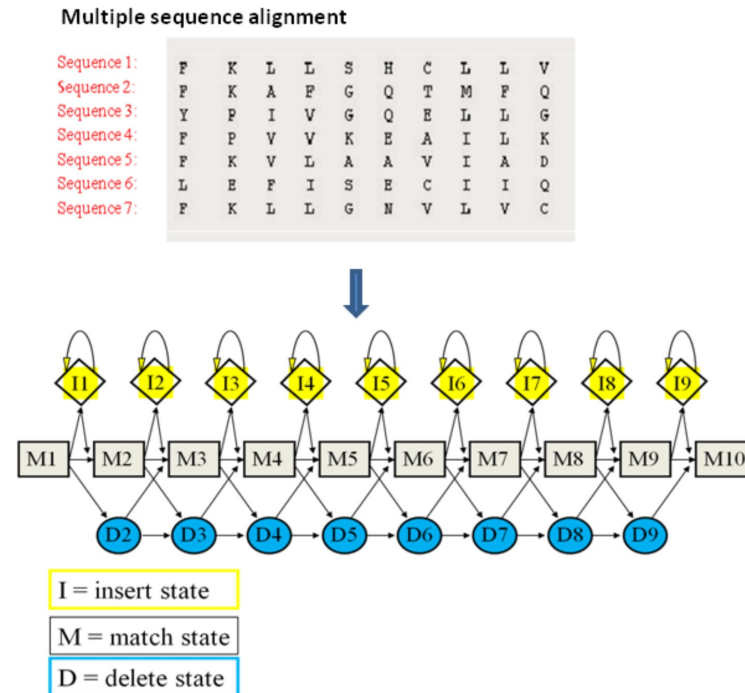


Figure 14 Representation of a Hidden Markov model based on a multiple sequence alignment. Amino acids are given a score at each position in the sequence alignment according to the frequency with which they occur. Transition probabilities (i.e., the likelihood that one particular amino acid follows another particular amino acid) and insertion and deletion states are also modelled.

InterPro is an integrative database

In InterPro, patterns, profiles, fingerprints and HMMs from a number of different databases are brought together into a single searchable resource, offering convenient access to their predictive capabilities without the need to visit the member databases individually.

By combining the different databases and signature types, InterPro capitalises on their individual strengths, producing a powerful tool for the prediction of protein function. InterPro aims to simplify and rationalise protein sequence analysis for the user by combining and organising information in a consistent manner, **removing redundancy, and adding extensive annotation** including GO terms and useful links about the signatures and the proteins they match.

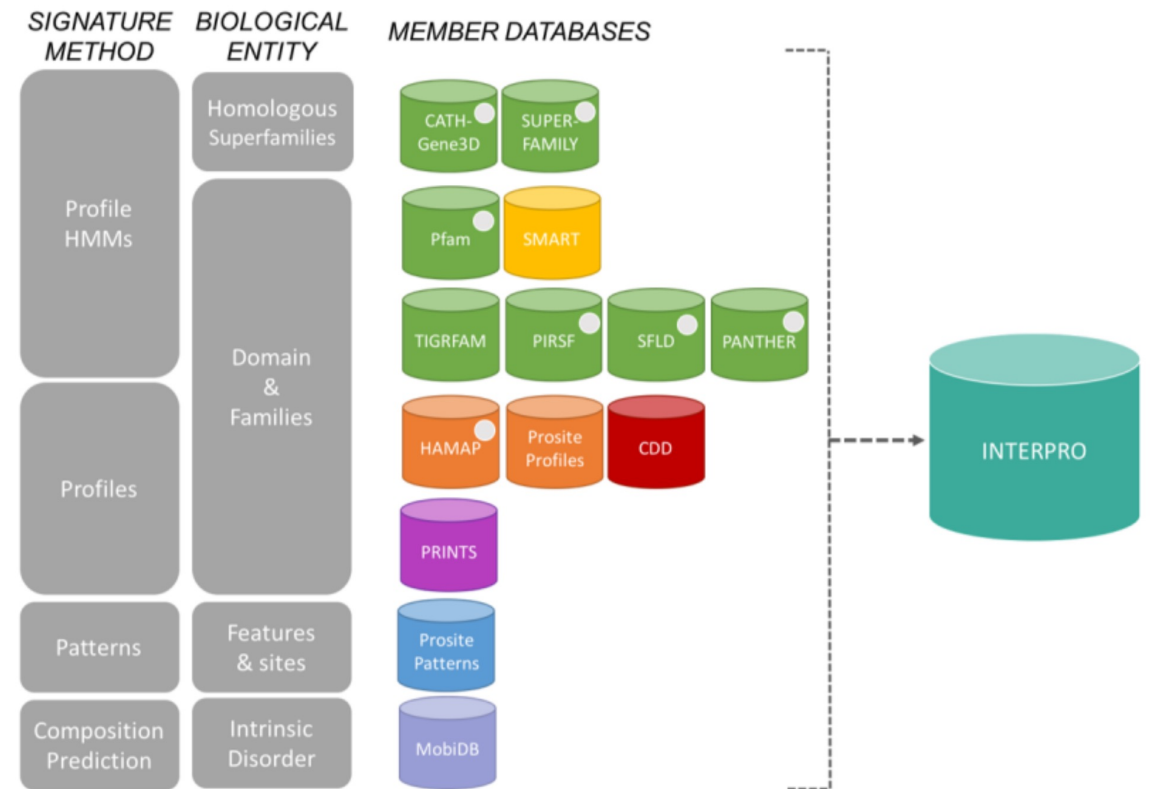


Figure 18 An overview of the different databases that are used to construct InterPro.

InterProSan can scan a query sequence against all the signatures aggregated in InterPro

InterPro Entry types

D Domain

Domains are distinct functional, structural or sequence units that may exist in a variety of biological contexts. A match to an InterPro entry of this type indicates the presence of a domain. Common examples of protein domains are the PH domain, Immunoglobulin domain or the classical C2H2 zinc finger.

F Family

A protein family is a group of proteins that share a common evolutionary origin reflected by their related functions, similarities in sequence, or similar primary, secondary or tertiary structure. A match to an InterPro entry of this type indicates membership of a protein family.

H Homologous Superfamily

A homologous superfamily is a group of proteins that share a common evolutionary origin, reflected by similarity in their structure. Since superfamily members often display very low similarity at the sequence level, this type of InterPro entry is usually based on a collection of underlying hidden Markov models, rather than a single signature. Homologous superfamilies usually comprise signatures from the SUPERFAMILY and CATH-Gene3D databases.

R Repeat

A short sequence that is typically repeated within a protein. Repeats are often relatively short <50 amino acids in length. Common repeats examples are Leucine Rich Repeats or WD40 repeats.

S Site

InterPro contains data for the following types of sites:

- **Active site** - A short sequence that contains one or more conserved residues, which allow the protein to bind to a ligand and carry out a catalytic activity.
- **Binding site** - A short sequence that contains one or more conserved residues, which form a protein interaction site.
- **Conserved site** - A short sequence that contains one or more conserved residues.
- **PTM site** - A short sequence that contains one or more conserved residues some of which are the site of a Post-translational modification.

U Unintegrated

InterPro 90.0 • 4th August 2022

Contents and coverage

InterPro protein matches are now calculated for all UniProtKB and UniParc proteins. InterPro release 90.0 contains [40 597](#) entries (last entry: [IPR046937](#)), representing:

H	Homologous Superfamily	3 399
F	Family	23 997
D	Domain	11 954
R	Repeat	327
S	Site	
	:.. Active Site	132
	:.. Binding Site	75
	:.. Conserved Site	696
	:.. PTM	17

Interpro cites 53778 publications in PubMed.

InterPro Entries are linked to GO terms

InterPro2GO

A total number of 35 086 GO terms mapped to InterPro entries.
These are available through the EBI GO browser QuickGO

The assignment of GO terms to InterPro entries is performed manually, and is an ongoing process

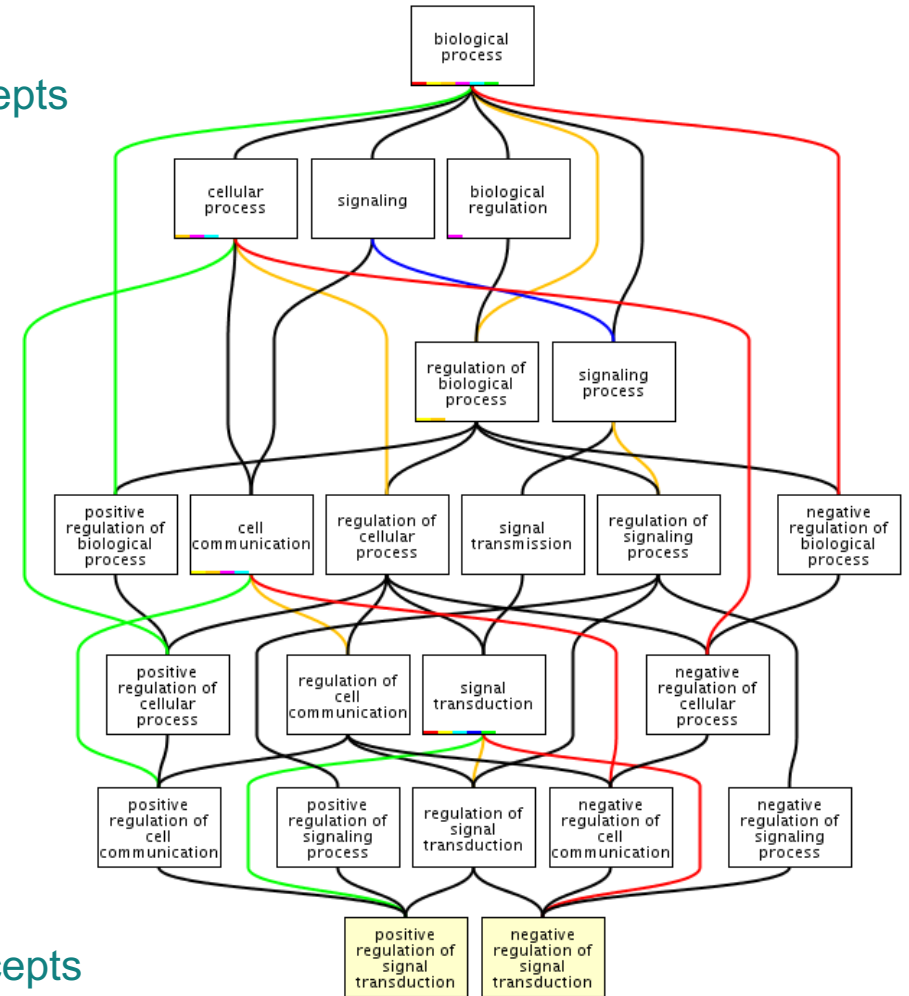
GO annotation in InterPro: why stability does not indicate accuracy in a sea of changing annotations

**Amaia Sangrador-Vegas^{1,*}, Alex L. Mitchell¹, Hsin-Yu Chang¹,
Siew-Yit Yong¹ and Robert D. Finn¹**

The Gene Ontology

- A way to capture biological knowledge for individual gene products in a written and computable form
- A set of concepts and their relationships to each other arranged as a hierarchy

Less specific concepts

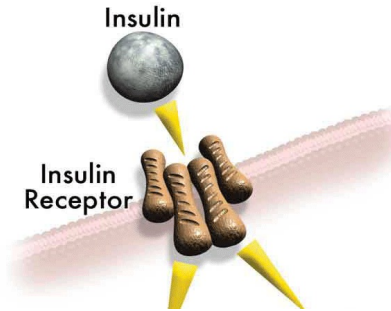


More specific concepts

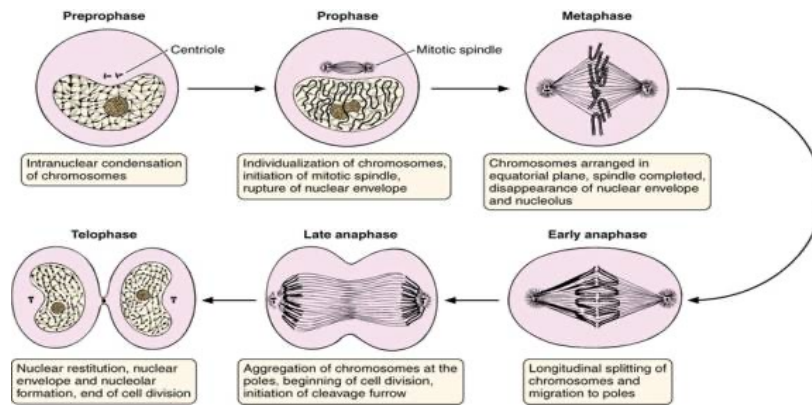
The domains in GO

1. Molecular Function

An elemental activity or task or job



- protein kinase activity
- insulin receptor activity



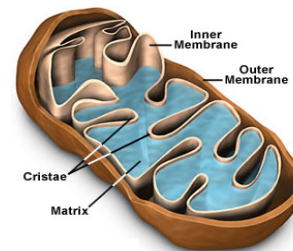
2. Biological Process

A commonly recognised series of events

- cell division

3. Cellular Component

Where a gene product is located



Anatomy of a GO term

i ID	GO:0005634	Unique identifier
i Name	nucleus	Term name
i Definition	A membrane-bounded organelle of eukaryotic cells in which chromosomes are housed and replicated. In most cells, the nucleus contains all of the cell's chromosomes except the organellar chromosomes, and is the site of RNA synthesis and processing. In some species, or in specialized cell types, RNA metabolism or DNA replication may be absent.	
i Comment		
i Synonyms		
Type	Synonym	Synonyms
exact	cell nucleus	
Cross-references associated with this term:		
Database	ID	Cross-references
INTERPRO	IPR000003	
INTERPRO	IPR000116	
INTERPRO	IPR000135	

GO evidence codes

Evidence codes fall into six general categories:

- experimental evidence

- phylogenetic evidence

- computational evidence

- author statements

- curatorial statements

- automatically generated annotations

- Inferred from Experiment (EXP)
- Inferred from Direct Assay (IDA)
- Inferred from Physical Interaction (IPI)
- Inferred from Mutant Phenotype (IMP)
- Inferred from Genetic Interaction (IGI)
- Inferred from Expression Pattern (IEP)
- Inferred from High Throughput Experiment (HTP)
- Inferred from High Throughput Direct Assay (HDA)
- Inferred from High Throughput Mutant Phenotype (HMP)
- Inferred from High Throughput Genetic Interaction (HGI)
- Inferred from High Throughput Expression Pattern (HEP)

What is Gene Ontology ?

Statistics for release 2022-09 ▾

Ontology

Property	Value
Valid terms	43335 ($\Delta = -223$)
Obsoleted terms	4008 ($\Delta = 238$)
Merged terms	2431 ($\Delta = 44$)
Biological process terms	28056
Molecular function terms	11242
Cellular component terms	4037

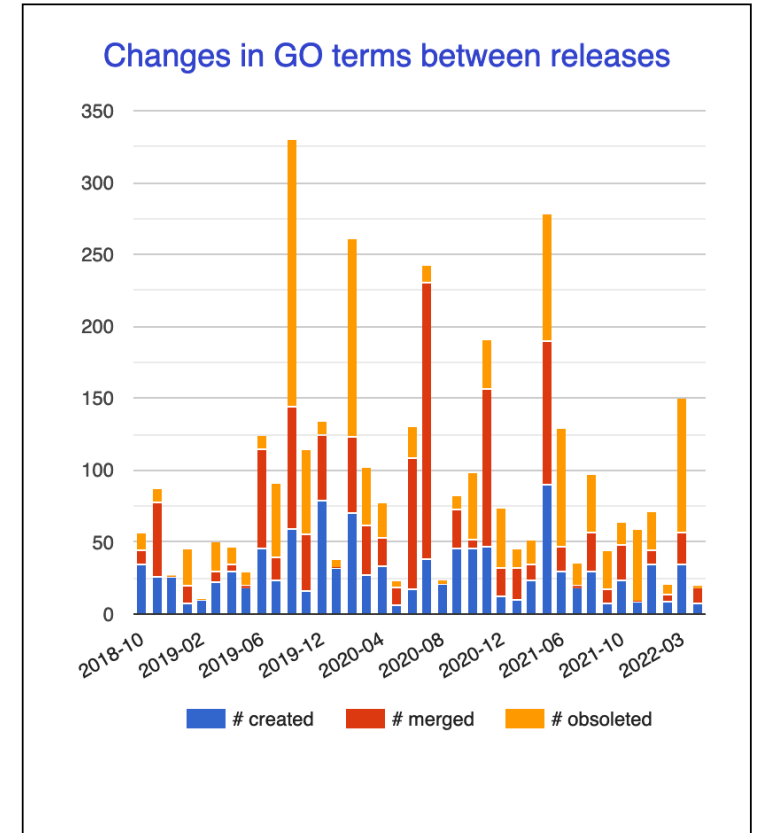
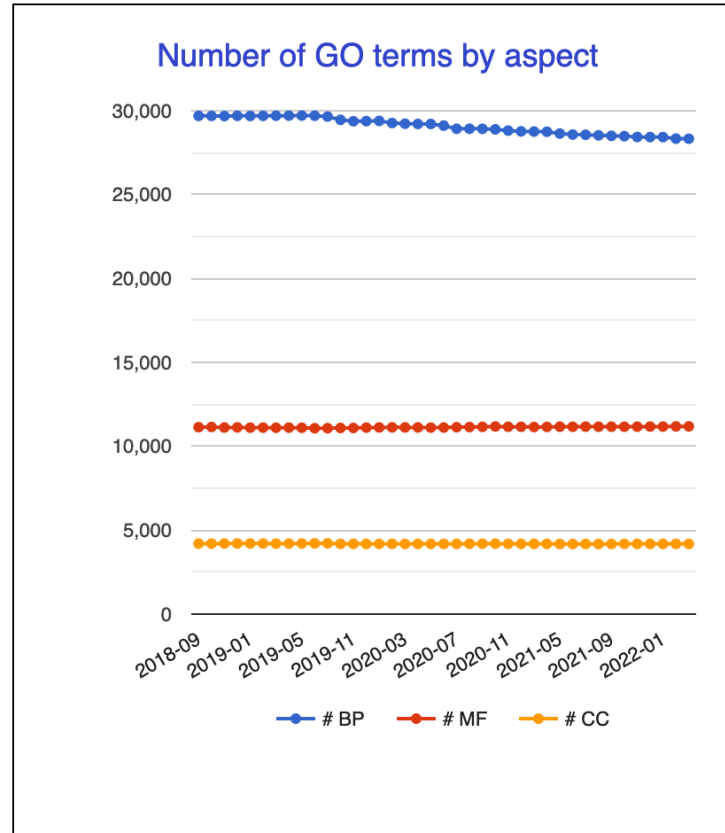
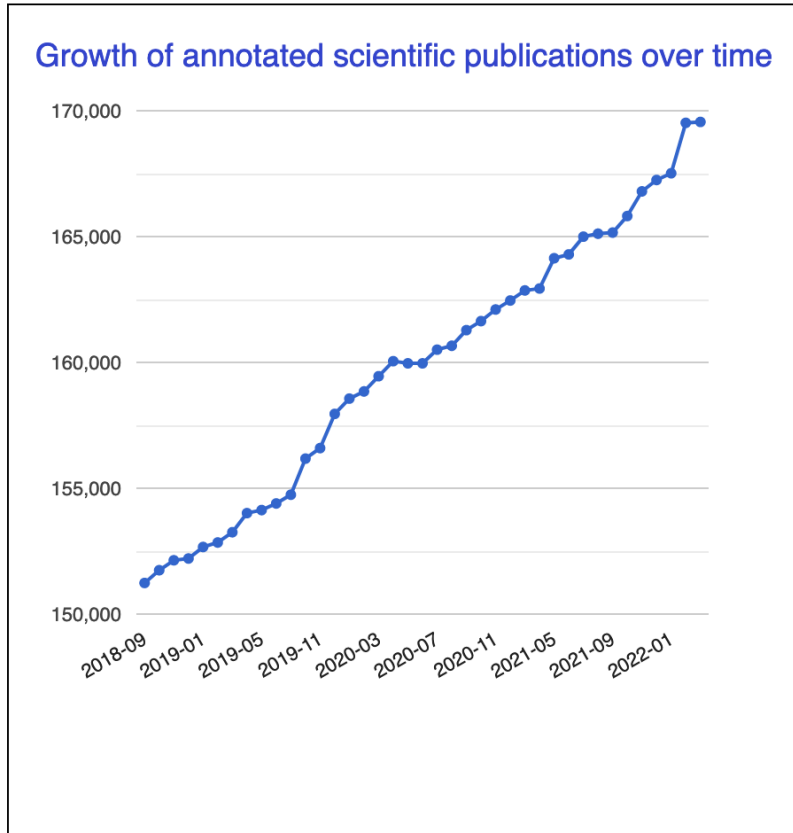
Annotations

Property	Value
Number of annotations	7,493,159
Annotations for biological process	2,831,298
Annotations for molecular function	2,362,896
Annotations for cellular component	2,298,965
Annotations for evidence PHYLO	3,993,128
Annotations for evidence IEA	1,385,272
Annotations for evidence OTHER	869,051
Annotations for evidence EXP	934,536
Annotations for evidence ND	252,132
Annotations for evidence HTP	59,040
Number of annotated scientific publications	172,379

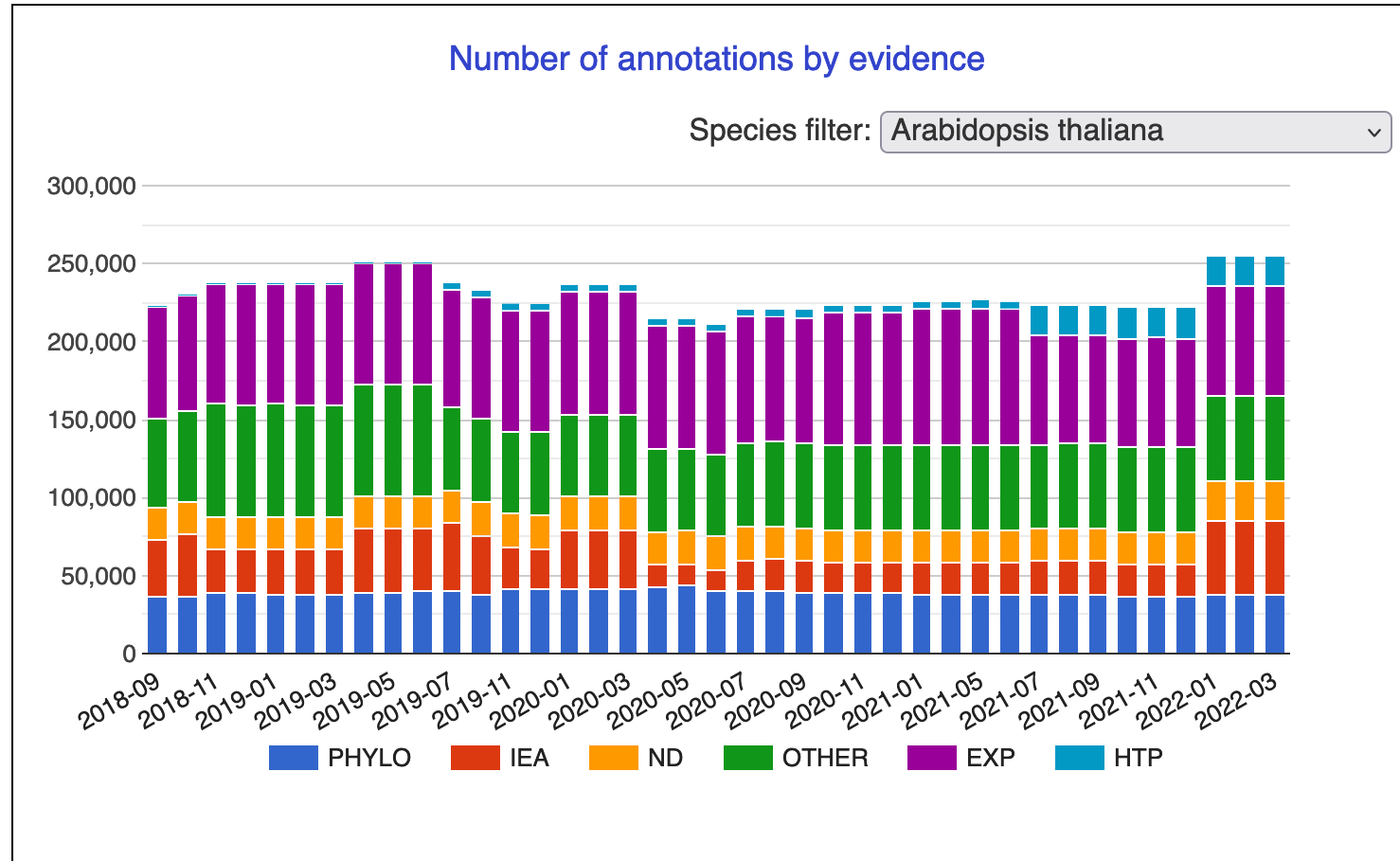
Gene products and species

Property	Value
Annotated gene products	1,483,687
Annotated species	5,257
Annotated species with over 1,000 annotations	185

What is Gene Ontology ?



What is Gene Ontology ?



Inferred from Experiment (EXP)

Inferred from Experiment (EXP)

Electronic (IEA) annotation are not manually reviewed

No biological Data available (ND)

What can we do with Gene Ontology ?

Over-representation analysis (ORA) in a nutshell

- Given:
 1. Gene list: e.g. RRP6, MRD1, RRP7, RRP43, RRP42 (yeast)
 2. Gene annotations: e.g. Gene ontology, transcription factor binding sites in promoter
- ORA Question: *Are any of the gene annotations surprisingly enriched in the gene list?*
- Details:
 - How to assess “surprisingly” (statistics)
 - How to correct for repeating the tests

What can we do with Gene Ontology ?

TopGO:

Some GO-endorsed enrichment tools are:

- [BiNGO](#)
- [GeneWeaver](#)
- [gProfiler](#)
- [GOzilla](#)
- [Ontologizer](#)

BIOINFORMATICS ORIGINAL PAPER Vol. 22 no. 13 2006, pages 1600–1607
doi:10.1093/bioinformatics/btl140

Gene expression

Improved scoring of functional groups from gene expression data by decorrelating GO graph structure

Adrian Alexa*, Jörg Rahnenführer and Thomas Lengauer

Max-Planck-Institute for Informatics, Stuhlsatzenhausweg 85, D-66123 Saarbrücken, Germany

Received on September 28, 2005; revised on March 30, 2006; accepted on April 4, 2006

Advance Access publication April 10, 2006

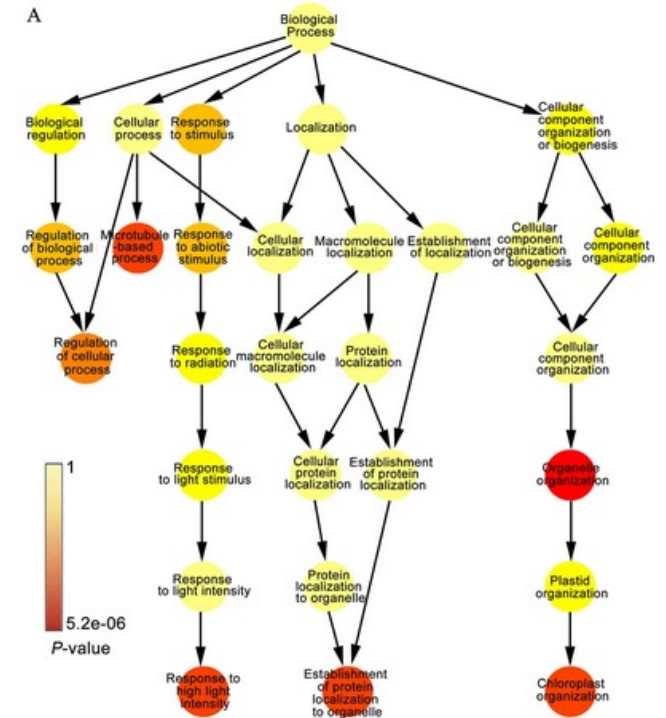
Associate Editor: Martin Bishop

Results:

We present two novel algorithms that improve GO group scoring **using the underlying GO graph topology**.

We show that both methods eliminate local dependencies between GO terms and point to relevant areas in the GO graph that remain undetected with state-of-the-art algorithms for scoring functional terms.

A simulation study demonstrates that the new methods exhibit a higher level of detecting relevant biological terms than competing methods



Let's have a look at InterProScan results

Using 4%

! UseGalaxy.fr will be undergoing maintenance from October 6th to 7th. Running jobs will be stopped. Thank you for your understanding.

Tools

✕
input type="text" value="eggnog"/>

eggnog Mapper functional sequence annotation by orthology

Funannotate functional annotation

WORKFLOWS

All workflows

FUN_001534-T1	cbc7bf376db3f01bccb47b8ecad0a96f	249	MobiDBLite	mobidb-lite	consensus disorder prediction
FUN_001534-T1	cbc7bf376db3f01bccb47b8ecad0a96f	249	MobiDBLite	mobidb-lite	consensus disorder prediction
FUN_001534-T1	cbc7bf376db3f01bccb47b8ecad0a96f	249	Gene3D	G3DSA:2.30.30.190	-
FUN_001534-T1	cbc7bf376db3f01bccb47b8ecad0a96f	249	MobiDBLite	mobidb-lite	consensus disorder prediction
FUN_001534-T1	cbc7bf376db3f01bccb47b8ecad0a96f	249	SUPERFAMILY	SSF74924	Cap-Gly domain
FUN_008567-T1	30f0e5f7e973401301eef93092cc7760	122	PANTHER	PTHR47640	TRNA SELENOCYSTEINE 1-ASSOCIATED P
FUN_008567-T1	30f0e5f7e973401301eef93092cc7760	122	CDD	cd12611	RRM1_NGR1_NAM8_like
FUN_008567-T1	30f0e5f7e973401301eef93092cc7760	122	SUPERFAMILY	SSF54928	RNA-binding domain, RBD
FUN_008567-T1	30f0e5f7e973401301eef93092cc7760	122	Pfam	PF00076	RNA recognition motif. (a.k.a. RRM, RBD, or
FUN_008567-T1	30f0e5f7e973401301eef93092cc7760	122	ProSiteProfiles	PS50102	Eukaryotic RNA Recognition Motif (RRM) pr
FUN_008567-T1	30f0e5f7e973401301eef93092cc7760	122	SMART	SM00360	rrm1_1
FUN_008567-T1	30f0e5f7e973401301eef93092cc7760	122	Gene3D	G3DSA:3.30.70.330	-
FUN_008567-T1	30f0e5f7e973401301eef93092cc7760	122	PANTHER	PTHR47640:SF10	TRNA SELENOCYSTEINE 1-ASSOCIATED P
FUN_001193-T1	fe673002d1a2aac7dcfa1d3b66bf6d2a	111	MobiDBLite	mobidb-lite	consensus disorder prediction
FUN_001193-T1	fe673002d1a2aac7dcfa1d3b66bf6d2a	111	Coils	Coil	Coil
FUN_003130-T1	bd4bab44ef23e2cf17f3e0d1c992e018	634	PANTHER	PTHR24012:SF466	POLYADENYLATE-BINDING PROTEIN 1-LIKI
FUN_003130-T1	bd4bab44ef23e2cf17f3e0d1c992e018	634	ProSiteProfiles	PS50102	Eukaryotic RNA Recognition Motif (RRM) pr
FUN_003130-T1	bd4bab44ef23e2cf17f3e0d1c992e018	634	Gene3D	G3DSA:3.30.70.330	-
FUN_003130-T1	bd4bab44ef23e2cf17f3e0d1c992e018	634	ProSiteProfiles	PS50102	Eukaryotic RNA Recognition Motif (RRM) pr
FUN_003130-T1	bd4bab44ef23e2cf17f3e0d1c992e018	634	TIGRFAM	TIGR01628	PABP-1234: polyadenylate binding protein,
FUN_003130-T1	bd4bab44ef23e2cf17f3e0d1c992e018	634	PANTHER	PTHR24012:SF466	POLYADENYLATE-BINDING PROTEIN 1-LIKI
FUN_003130-T1	bd4bab44ef23e2cf17f3e0d1c992e018	634	Pfam	PF00076	RNA recognition motif. (a.k.a. RRM, RBD, or
FUN_003130-T1	bd4bab44ef23e2cf17f3e0d1c992e018	634	Pfam	PF00076	RNA recognition motif. (a.k.a. RRM, RBD, or

History

MucorProtSet

36 shown, 73 deleted, 1881 hidden

3.99 GB

1903: cd-hit on data 3: Clusters

~140,000 lines

format: **tabular**, génome de référence: ?

21/09/2022 08:48:42:801 Welcome to InterProScan-5.55-88.0

21/09/2022 08:48:42:803 Running InterProScan v5 in STANDALONE mode... on Linux

21/09/2022 08:48:49:153 RunID: cpu-node-18.ifb.local_20220921_0848488

21/09/2022 08:48:56:739 Loading file /sh

1902: InterProScan on d ata 3 (tsv)

Let's have a look at InterProScan results

The TSV format presents the match data in columns as follows:

- Protein Accession (e.g. P51587)
- Sequence MD5 digest (e.g. 14086411a2cdf1c4cba63020e1622579)
- Sequence Length (e.g. 3418)
- Analysis (e.g. Pfam / PRINTS / Gene3D)
- Signature Accession (e.g. PF09103 / G3DSA:2.40.50.140)
- Signature Description (e.g. BRCA2 repeat profile)
- Start location
- Stop location
- Score - is the e-value of the match reported by member database method (e.g. 3.1E-52)
- Status - is the status of the match (T: true)
- Date - is the date of the run
- (InterProScan annotations - accession (e.g. IPR002093) - optional column; only displayed if -iprscan option is switched on)
- (InterProScan annotations - description (e.g. BRCA2 repeat) - optional column; only displayed if -iprscan option is switched on)
- (GO annotations (e.g. [GO:0005515](#)) - optional column; only displayed if --goterms option is switched on)
- (Pathways annotations (e.g. REACT_71) - optional column; only displayed if --pathways option is switched on)

What's the fraction of annotated sequences ?

Galaxy France Workflow Visualize Données partagées Aide Utilisateur

! UseGalaxy.fr will be undergoing maintenance from October 6th to 7th. Running jobs will be stopped. Thank you for your understanding.

Tools

Sort

Upload Data

Show Sections

EXTRACT FEATURES FROM GFF data

Sub-sample sequences files e.g. to reduce coverage

Filter sequences by ID from a tabular file

Column arrange by header name

VCFsort: Sort VCF dataset by coordinate

obisort sorts sequence records according to the value of a given attribute

SortSam sort SAM/BAM dataset

Sort.seqs put sequences in different files in the same order

Sort assembly

VSearch sorting

Sort data in ascending or descending order

WORKFLOWS

All workflows

Sort data in ascending or descending order (Galaxy Version 1.1.1)

Sort Query

1988: Advanced Cut on data 1902

Number of header lines

0

These will be ignored during sort.

Column selections

1: Column selections

on column

Column: 1

in

Ascending order

Descending order

Flavor

Fast numeric sort (-n)

General numeric sort (scientific notation -g)

Natural/Version sort (-V)

Alphabetical sort

Human-readable numbers (-h)

Random order (-R)

+ Insert Column selections

Output unique values

No

What's the fraction of annotated sequences ?

Galaxy France

Workflow Visualize Données partagées Aide Utilisateur

! UseGalaxy.fr will be undergoing maintenance from October 6th to 7th. Running jobs will be stopped. Thank you for your understanding.

Tools

Sort

Upload Data

Show Sections

order

Sort Column Order by heading

Filter data on any column using simple expressions

Select lines that match an expression

Filter GTF data by attribute values_list

Filter GFF data by attribute using simple expressions

Filter GFF data by feature count using simple expressions

Extract features from GFF data

Sub-sample sequences files e.g. to reduce coverage

Filter sequences by ID from a tabular file

Column arrange by header name

VCFSort: Sort VCF dataset by coordinate

obisort sorts sequence records according to the value of a given attribute

SortSam sort SAM/BAM dataset

Sort.seqs put sequences in different files in the same order

Sort assembly

VSearch sorting

Sort data in ascending or descending order

WORKFLOWS

All workflows

Sort data in ascending or descending order (Galaxy Version 1.1.1)

Sort Query

1984: InterProScan on data 3 (tsv)

Number of header lines

0

These will be ignored during sort.

Column selections

1: Column selections

on column

Column: 1

in

Ascending order

Descending order

Flavor

Fast numeric sort (-n)

General numeric sort (scientific notation -g)

Natural/Version sort (-V)

Alphabetical sort

Human-readable numbers (-h)

Random order (-R)

+ Insert Column selections

Output unique values

Yes

Print only unique values, based on sorted key columns. See help section for details. (--unique)

Ignore case

No

Sort and Join key column values regardless of upper/lower case letters. (-i)

Email notification

No

Send an email notification when the job completes.

Execute

What's the fraction of annotated sequences ?

Galaxy France

Workflow Visualize Données partagées Aide Utilisateur Using 11%

! UseGalaxy.fr will be undergoing maintenance from October 6th to 7th. Running jobs will be stopped. Thank you for your understanding.

Tools

Sort

Upload Data

Show Sections

order

Sort Column Order by heading

Filter data on any column using simple expressions

Select lines that match an expression

Filter GTF data by attribute values_list

Filter GFF data by attribute using simple expressions

Accession	Sequence ID	Length	Database	Tool	Annotation
FUN_000001-T1	1f9cb956ae0dcbdbd1b334d9b10ec0e9	227	MobiDBLite	mobidb-lite	consensus disorder prediction
FUN_000002-T1	2d68be6c38afea689ce7d6783b888576	189	Coils	Coil	Coil
FUN_000005-T1	47cc622a1d31a46d5070a14eab3796f0	322	SUPERFAMILY	SSF53474	alpha/beta-Hydrolases
FUN_000006-T1	4b7494c2d2ab8ffe9e2047c26ea36e41	647	PANTHER	PTHR19879	TRANSCRIPTION INITIATION FACTOR
FUN_000007-T1	adc95b912482bc4661ea331f7c0e8336	299	Pfam	PF08731	Transcription factor AFT
FUN_000008-T1	d77daa28212e457b390a9e6fd807af09	723	MobiDBLite	mobidb-lite	consensus disorder prediction
FUN_000009-T1	e37b2faf08d18b292e46c4dff4eefb1f	137	CDD	cd00923	Cyt_c_Oxidase_Va
FUN_000010-T1	348b0e908f1352d758b5ba19825dc6bc	521	PANTHER	PTHR11606:SF13	GLUTAMATE DEHYDROGENASE 1, M
FUN_000013-T1	b2c3abaaa8f28c8b6690e080b5d8e70f	1050	MobiDBLite	mobidb-lite	consensus disorder prediction
FUN_000014-T1	b832b7036c881a1382f230f0a732d757	391	Gene3D	G3DSA:2.40.30.10	Translation factors
FUN_000015-T1	ad423c24eb3830bf3cbff1d9fd1e1449	582	Coils	Coil	Coil
FUN_000016-T1	3a7fbb5f37cd628b22b9cd30e287de7f	370	Gene3D	G3DSA:3.80.10.10	Ribonuclease Inhibitor
FUN_000017-T1	e1e0b10fe59ca83ff5684bce0ed36343	266	Hamap	MF_03122	40S ribosomal protein S1 [RPS3A].
FUN_000018-T1	012afa75b0424eef251ba33dd960ea07	147	PANTHER	PTHR23050	CALCIUM BINDING PROTEIN
FUN_000019-T1	8ba4ca76f7a7fa43d9b20b7fc586e6e7	427	CDD	cd00610	OAT_like
FUN_000020-T1	10874220baee79ba95e2703bde57221e	661	PANTHER	PTHR10073:SF12	DNA MISMATCH REPAIR PROTEIN M

History

Rechercher des données

MucorProtSet

38 shown, 73 deleted, 1881 hidden

11.45 GB

1989: Sort on data 1984

11,346 lines

format: **tabular**, génome de référence: ?

Galaxy France

Workflow Visualize Données partagées Aide Utilisateur Using 11%

! UseGalaxy.fr will be undergoing maintenance from October 6th to 7th. Running jobs will be stopped. Thank you for your understanding.

Tools

Sort

Upload Data

Show Sections

order

Sort Column Order by heading

Filter data on any column using simple expressions

Select lines that match an expression

Filter GTF data by attribute values_list

This dataset is large and only the first megabyte is shown below. Show all | Save

```
>FUN_000001-T1 FUN_000001
MFSGSSSNKNEGVPKSKPGKVGSKRRDPHAARS GPKRMMNQPNNLQRLDQMSLFR TGSGREQDQSMSTFSYGNVGMSTMH
SPVAEDIRNRQMNNDVPIVVESEQNVEGTESDEEEMVGNVTNISDDYANTAAEQELTEEFEEIIDEQVGNSTPEDSFV
GKYVTEIQNRLKGGVKPIEYQRKTYWVNVVEYEGFDYNTRVNPDIFRPRVFIWLPDILNNGGANDQF
>FUN_000002-T1 FUN_000002
MICPSESQRKLLDGI GFSKGVESVMMECSGEEDGNHTEEDILKLEMYTSNCLKNEINQYQASWTTSGRRRIFAICQI
GNKMLLSTSRMGIGKWCFFVQIRSAI VPRDWDREYRLNRMVLLMKL KELLLEQEEVMTMLKQQQSRQSPVESMDQAGKK
ILANHTQVKVANHVNGDIDEGFSTNQKKI
>FUN_000003-T1 FUN_000003
MKVNTIVKSYCYGIAPELLHAKELHRLYQRGDDLSTEQL EARSKGLSCILDLEPETEGTLRCLFSEDLWTKLTAKYTL
RFKTA PSAIDMSLIEKWSYIMNLYDQQNNVRAKRYLNQLKSDQNIKDVNEKVFDFYEEILILASEDAVRQSQENGEP
>FUN_000004-T1 FUN_000004
MLDTRNASKISERDYIYQIWLPLL SKLYNINKNIVRIKTGETYSENTTESKANLYNHNTHIIGFKDRLRILVDFDDEEFD
LVCGEGLRDASDKKISSDISKLAREGKEAEVAIQIYNSMDKYSIKSRRAWLQCFIGPRCIFSTIHATKHQYHVMPEFS
I TEPTSEI GSDGTSSTRPI ETERDSVEKAAI ATKETI GENKOKTTAKNSRRI SFTPEKI NTPPEPTWFTPHADRSI S
```

History

Rechercher des données

MucorProtSet

38 shown, 73 deleted, 1881 hidden

11.45 GB

3: Galaxy13-[Funannotat e_predict_annotation_o n_data_4,_data_9,_and_data_6__pr otin_sequences].fasta

13,403 sequences

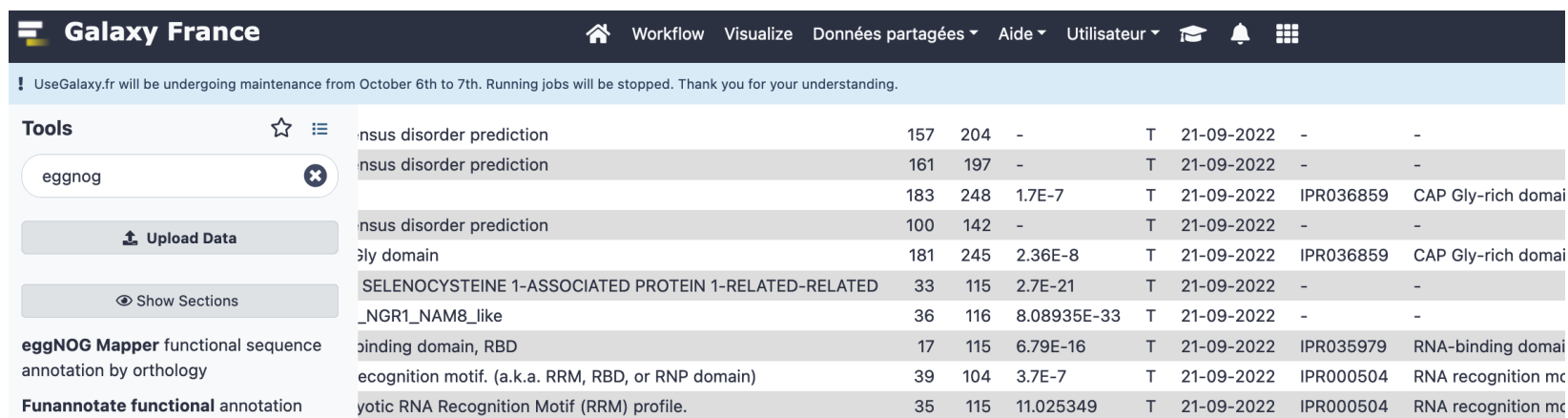
format: **fasta**, génome de référence: ?

84% are annotated

How reliable is an InterProScan Hit ?

Why are there no e-values associated with InterPro entries?

The signatures contained within InterPro are produced in different ways by different member databases, so their e-values and/or scoring systems cannot be meaningfully compared or combined. For this reason, we do not show e-values on the InterPro web site. However, e-values can be obtained via the downloadable InterProScan software package, which outputs detailed individual results for each member database sequence analysis algorithm.



Tool	Description	Hit 1	Hit 2	E-value	Database	Accession	Annotation
eggNOG	Insus disorder prediction	157	204	-	T	21-09-2022	-
	Insus disorder prediction	161	197	-	T	21-09-2022	-
	Insus disorder prediction	183	248	1.7E-7	T	21-09-2022	IPR036859 CAP Gly-rich domain
	Insus disorder prediction	100	142	-	T	21-09-2022	-
	Gly domain	181	245	2.36E-8	T	21-09-2022	IPR036859 CAP Gly-rich domain
	SELENOCYSTEINE 1-ASSOCIATED PROTEIN 1-RELATED-RELATED	33	115	2.7E-21	T	21-09-2022	-
	_NGR1_NAM8_like	36	116	8.08935E-33	T	21-09-2022	-
	Binding domain, RBD	17	115	6.79E-16	T	21-09-2022	IPR035979 RNA-binding domain
	Recognition motif. (a.k.a. RRM, RBD, or RNP domain)	39	104	3.7E-7	T	21-09-2022	IPR000504 RNA recognition motif
	Cytosolic RNA Recognition Motif (RRM) profile.	35	115	11.025349	T	21-09-2022	IPR000504 RNA recognition motif

The E-value (expectation value) is the number of hits that would be expected to have a score equal to or better than this value, by chance alone. This means that a good E-value which gives a confident prediction is much less than 1. E-values around 1 is what is expected by chance. Thus, the lower the E-value, the more specific the search for domains will be.

How reliable is an InterProScan Hit ?

FUN_003130-T1	PANTHER	PTHR24012:SF466	2.4E-205	
FUN_003130-T1	ProSiteProfiles	PS50102	16.807129	MetaCyc: PWY-102 MetaCyc: PWY-1061 MetaCyc: PWY-1187 MetaCyc: PWY-2083 MetaCyc: PWY-3542 MetaCyc: PWY-4021 MetaCyc: PWY-4161 MetaCyc: PWY-4202 MetaCyc: PWY-5035 MetaCyc: PWY-5036 MetaCyc:
FUN_003130-T1	Gene3D	G3DSA:3.30.70.330	4.0E-25	MetaCyc: PWY-102 MetaCyc: PWY-1061 MetaCyc: PWY-1187 MetaCyc: PWY-2083 MetaCyc: PWY-3542 MetaCyc: PWY-4021 MetaCyc: PWY-4161 MetaCyc: PWY-4202 MetaCyc: PWY-5035 MetaCyc: PWY-5036 MetaCyc:
FUN_003130-T1	ProSiteProfiles	PS50102	18.137451	MetaCyc: PWY-102 MetaCyc: PWY-1061 MetaCyc: PWY-1187 MetaCyc: PWY-2083 MetaCyc: PWY-3542 MetaCyc: PWY-4021 MetaCyc: PWY-4161 MetaCyc: PWY-4202 MetaCyc: PWY-5035 MetaCyc: PWY-5036 MetaCyc:
FUN_003130-T1	TIGRFAM	TIGR01628	6.1E-222	Reactome: R-DDI-156827 Reactome: R-DDI-975956 Reactome: R-DDI-975957 Reactome: R-DME-156827 Reactome: R-DME-429947 Reactome: R-DME-450408 Reactome: R-DME-72649 Reactome: R-DME-975956 Rea
FUN_003130-T1	PANTHER	PTHR24012:SF466	2.4E-205	
FUN_003130-T1	Pfam	PF00076	2.0E-22	MetaCyc: PWY-102 MetaCyc: PWY-1061 MetaCyc: PWY-1187 MetaCyc: PWY-2083 MetaCyc: PWY-3542 MetaCyc: PWY-4021 MetaCyc: PWY-4161 MetaCyc: PWY-4202 MetaCyc: PWY-5035 MetaCyc: PWY-5036 MetaCyc:
FUN_003130-T1	Pfam	PF00076	1.1E-20	MetaCyc: PWY-102 MetaCyc: PWY-1061 MetaCyc: PWY-1187 MetaCyc: PWY-2083 MetaCyc: PWY-3542 MetaCyc: PWY-4021 MetaCyc: PWY-4161 MetaCyc: PWY-4202 MetaCyc: PWY-5035 MetaCyc: PWY-5036 MetaCyc:
FUN_003130-T1	Pfam	PF00076	1.8E-18	MetaCyc: PWY-102 MetaCyc: PWY-1061 MetaCyc: PWY-1187 MetaCyc: PWY-2083 MetaCyc: PWY-3542 MetaCyc: PWY-4021 MetaCyc: PWY-4161 MetaCyc: PWY-4202 MetaCyc: PWY-5035 MetaCyc: PWY-5036 MetaCyc:
FUN_003130-T1	Pfam	PF00076	3.1E-20	MetaCyc: PWY-102 MetaCyc: PWY-1061 MetaCyc: PWY-1187 MetaCyc: PWY-2083 MetaCyc: PWY-3542 MetaCyc: PWY-4021 MetaCyc: PWY-4161 MetaCyc: PWY-4202 MetaCyc: PWY-5035 MetaCyc: PWY-5036 MetaCyc:
FUN_003130-T1	Gene3D	G3DSA:3.30.70.330	4.2E-29	MetaCyc: PWY-102 MetaCyc: PWY-1061 MetaCyc: PWY-1187 MetaCyc: PWY-2083 MetaCyc: PWY-3542 MetaCyc: PWY-4021 MetaCyc: PWY-4161 MetaCyc: PWY-4202 MetaCyc: PWY-5035 MetaCyc: PWY-5036 MetaCyc:
FUN_003130-T1	Gene3D	G3DSA:3.30.70.330	1.9E-30	MetaCyc: PWY-102 MetaCyc: PWY-1061 MetaCyc: PWY-1187 MetaCyc: PWY-2083 MetaCyc: PWY-3542 MetaCyc: PWY-4021 MetaCyc: PWY-4161 MetaCyc: PWY-4202 MetaCyc: PWY-5035 MetaCyc: PWY-5036 MetaCyc:
FUN_003130-T1	Gene3D	G3DSA:1.10.1900.10	1.3E-34	
FUN_003130-T1	CDD	cd12381	1.08824E-45	
FUN_003130-T1	Coils	Coil	-	
FUN_003130-T1	CDD	cd12379	4.48837E-53	Reactome: R-DDI-156827 Reactome: R-DDI-975956 Reactome: R-DDI-975957 Reactome: R-DME-156827 Reactome: R-DME-429947 Reactome: R-DME-450408 Reactome: R-DME-72649 Reactome: R-DME-975956 Rea
FUN_003130-T1	Coils	Coil	-	
FUN_003130-T1	CDD	cd12380	1.37304E-41	
FUN_003130-T1	MobiDBLite	mobidb-lite	-	
FUN_003130-T1	MobiDBLite	mobidb-lite	-	
FUN_003130-T1	MobiDBLite	mobidb-lite	-	
FUN_003130-T1	Gene3D	G3DSA:3.30.70.330	5.3E-31	MetaCyc: PWY-102 MetaCyc: PWY-1061 MetaCyc: PWY-1187 MetaCyc: PWY-2083 MetaCyc: PWY-3542 MetaCyc: PWY-4021 MetaCyc: PWY-4161 MetaCyc: PWY-4202 MetaCyc: PWY-5035 MetaCyc: PWY-5036 MetaCyc:
FUN_003130-T1	MobiDBLite	mobidb-lite	-	
FUN_003130-T1	ProSiteProfiles	PS50102	20.576372	MetaCyc: PWY-102 MetaCyc: PWY-1061 MetaCyc: PWY-1187 MetaCyc: PWY-2083 MetaCyc: PWY-3542 MetaCyc: PWY-4021 MetaCyc: PWY-4161 MetaCyc: PWY-4202 MetaCyc: PWY-5035 MetaCyc: PWY-5036 MetaCyc:
FUN_003130-T1	SUPERFAMILY	SSF54928	3.33E-54	MetaCyc: PWY-102 MetaCyc: PWY-1061 MetaCyc: PWY-1187 MetaCyc: PWY-2083 MetaCyc: PWY-3542 MetaCyc: PWY-361 MetaCyc: PWY-4021 MetaCyc: PWY-4161 MetaCyc: PWY-4202 MetaCyc: PWY-4801 MetaCyc: P
FUN_003130-T1	Pfam	PF00658	9.0E-30	MetaCyc: PWY-7511 Reactome: R-DDI-156827 Reactome: R-DDI-975956 Reactome: R-DDI-975957 Reactome: R-DME-156827 Reactome: R-DME-429947 Reactome: R-DME-450408 Reactome: R-DME-72649 Reactome:
FUN_003130-T1	MobiDBLite	mobidb-lite	-	
FUN_003130-T1	SUPERFAMILY	SSF63570	1.27E-32	MetaCyc: PWY-7511 Reactome: R-DDI-156827 Reactome: R-DDI-975956 Reactome: R-DDI-975957 Reactome: R-DME-156827 Reactome: R-DME-429947 Reactome: R-DME-450408 Reactome: R-DME-72649 Reactome:
FUN_003130-T1	SMART	SM00361	2.1	MetaCyc: PWY-7511 Reactome: R-CEL-72163 Reactome: R-DDI-156827 Reactome: R-DDI-975956 Reactome: R-DDI-975957 Reactome: R-DME-159236 Reactome: R-DME-6781823 Reactome: R-DME-6782135 Reactome:
FUN_003130-T1	SMART	SM00361	0.48	MetaCyc: PWY-7511 Reactome: R-CEL-72163 Reactome: R-DDI-156827 Reactome: R-DDI-975956 Reactome: R-DDI-975957 Reactome: R-DME-159236 Reactome: R-DME-6781823 Reactome: R-DME-6782135 Reactome:
FUN_003130-T1	SMART	SM00361	0.33	MetaCyc: PWY-7511 Reactome: R-CEL-72163 Reactome: R-DDI-156827 Reactome: R-DDI-975956 Reactome: R-DDI-975957 Reactome: R-DME-159236 Reactome: R-DME-6781823 Reactome: R-DME-6782135 Reactome:
FUN_003130-T1	SMART	SM00361	0.2	MetaCyc: PWY-7511 Reactome: R-CEL-72163 Reactome: R-DDI-156827 Reactome: R-DDI-975956 Reactome: R-DDI-975957 Reactome: R-DME-159236 Reactome: R-DME-6781823 Reactome: R-DME-6782135 Reactome:
FUN_003130-T1	SUPERFAMILY	SSF54928	2.45E-49	MetaCyc: PWY-102 MetaCyc: PWY-1061 MetaCyc: PWY-1187 MetaCyc: PWY-2083 MetaCyc: PWY-3542 MetaCyc: PWY-361 MetaCyc: PWY-4021 MetaCyc: PWY-4161 MetaCyc: PWY-4202 MetaCyc: PWY-4801 MetaCyc: P
FUN_003130-T1	SMART	SM00517	1.1E-36	MetaCyc: PWY-7511 Reactome: R-DDI-156827 Reactome: R-DDI-975956 Reactome: R-DDI-975957 Reactome: R-DME-156827 Reactome: R-DME-429947 Reactome: R-DME-450408 Reactome: R-DME-72649 Reactome:
FUN_003130-T1	SMART	SM00360	3.2E-22	MetaCyc: PWY-102 MetaCyc: PWY-1061 MetaCyc: PWY-1187 MetaCyc: PWY-2083 MetaCyc: PWY-3542 MetaCyc: PWY-4021 MetaCyc: PWY-4161 MetaCyc: PWY-4202 MetaCyc: PWY-5035 MetaCyc: PWY-5036 MetaCyc:
FUN_003130-T1	SMART	SM00360	2.1E-25	MetaCyc: PWY-102 MetaCyc: PWY-1061 MetaCyc: PWY-1187 MetaCyc: PWY-2083 MetaCyc: PWY-3542 MetaCyc: PWY-4021 MetaCyc: PWY-4161 MetaCyc: PWY-4202 MetaCyc: PWY-5035 MetaCyc: PWY-5036 MetaCyc:
FUN_003130-T1	SMART	SM00360	1.3E-27	MetaCyc: PWY-102 MetaCyc: PWY-1061 MetaCyc: PWY-1187 MetaCyc: PWY-2083 MetaCyc: PWY-3542 MetaCyc: PWY-4021 MetaCyc: PWY-4161 MetaCyc: PWY-4202 MetaCyc: PWY-5035 MetaCyc: PWY-5036 MetaCyc:
FUN_003130-T1	SMART	SM00360	3.0E-28	MetaCyc: PWY-102 MetaCyc: PWY-1061 MetaCyc: PWY-1187 MetaCyc: PWY-2083 MetaCyc: PWY-3542 MetaCyc: PWY-4021 MetaCyc: PWY-4161 MetaCyc: PWY-4202 MetaCyc: PWY-5035 MetaCyc: PWY-5036 MetaCyc:
FUN_003130-T1	ProSiteProfiles	PS50102	19.757713	MetaCyc: PWY-102 MetaCyc: PWY-1061 MetaCyc: PWY-1187 MetaCyc: PWY-2083 MetaCyc: PWY-3542 MetaCyc: PWY-4021 MetaCyc: PWY-4161 MetaCyc: PWY-4202 MetaCyc: PWY-5035 MetaCyc: PWY-5036 MetaCyc:
FUN_003130-T1	Coils	Coil	-	
FUN_003130-T1	PANTHER	PTHR24012	2.4E-205	
FUN_003130-T1	SUPERFAMILY	SSF81995	2.17E-6	
FUN_003130-T1	CDD	cd12378	3.79251E-50	
FUN_003130-T1	ProSiteProfiles	PS51309	21.821808	MetaCyc: PWY-7511 Reactome: R-DDI-156827 Reactome: R-DDI-975956 Reactome: R-DDI-975957 Reactome: R-DME-156827 Reactome: R-DME-429947 Reactome: R-DME-450408 Reactome: R-DME-72649 Reactome:
FUN_003130-T1	PANTHER	PTHR24012	2.4E-205	

How reliable is an InterProScan Hit ?

Galaxy France Workflow Visualize Données partagées Aide Utilisateur

! UseGalaxy.fr will be undergoing maintenance from October 6th to 7th. Running jobs will be stopped. Thank you for your understanding.

Tools ☆ ☰

eggnog ✖

Upload Data

Show Sections

eggnOG Mapper functional sequence annotation by orthology

Funannotate functional annotation

WORKFLOWS

All workflows

nsus disorder prediction	157	204	-	T	21-09-2022	-	-
nsus disorder prediction	161	197	-	T	21-09-2022	-	-
nsus disorder prediction	183	248	1.7E-7	T	21-09-2022	IPR036859	CAP Gly-rich domai
nsus disorder prediction	100	142	-	T	21-09-2022	-	-
gly domain	181	245	2.36E-8	T	21-09-2022	IPR036859	CAP Gly-rich domai
SELENOCYSTEINE 1-ASSOCIATED PROTEIN 1-RELATED-RELATED	33	115	2.7E-21	T	21-09-2022	-	-
_NGR1_NAM8_like	36	116	8.08935E-33	T	21-09-2022	-	-
binding domain, RBD	17	115	6.79E-16	T	21-09-2022	IPR035979	RNA-binding domai
ecognition motif. (a.k.a. RRM, RBD, or RNP domain)	39	104	3.7E-7	T	21-09-2022	IPR000504	RNA recognition mc
yotic RNA Recognition Motif (RRM) profile.	35	115	11.025349	T	21-09-2022	IPR000504	RNA recognition mc
	36	109	2.3E-6	T	21-09-2022	IPR000504	RNA recognition mc
	34	114	9.7E-18	T	21-09-2022	IPR012677	Nucleotide-binding
SELENOCYSTEINE 1-ASSOCIATED PROTEIN 1-RELATED	33	115	2.7E-21	T	21-09-2022	-	-
nsus disorder prediction	73	111	-	T	21-09-2022	-	-
	7	27	-	T	21-09-2022	-	-
ADENYLATE-BINDING PROTEIN 1-LIKE	49	494	2.4E-205	T	21-09-2022	-	-
yotic RNA Recognition Motif (RRM) profile.	52	130	16.807129	T	21-09-2022	IPR000504	RNA recognition mc
	129	221	4.0E-25	T	21-09-2022	IPR012677	Nucleotide-binding
yotic RNA Recognition Motif (RRM) profile.	140	217	18.137451	T	21-09-2022	IPR000504	RNA recognition mc
-1234: polyadenylate binding protein, human types 1, 2, 3, 4 family	52	625	6.1E-222	T	21-09-2022	IPR006515	Polyadenylate bindi
ADENYLATE-BINDING PROTEIN 1-LIKE	546	628	2.4E-205	T	21-09-2022	-	-
ecognition motif. (a.k.a. RRM, RBD, or RNP domain)	235	303	2.0E-22	T	21-09-2022	IPR000504	RNA recognition mc
ecognition motif. (a.k.a. RRM, RBD, or RNP domain)	142	210	1.1E-20	T	21-09-2022	IPR000504	RNA recognition mc
ecognition motif. (a.k.a. RRM, RBD, or RNP domain)	54	123	1.8E-18	T	21-09-2022	IPR000504	RNA recognition mc
ecognition motif. (a.k.a. RRM, RBD, or RNP domain)	338	407	3.1E-20	T	21-09-2022	IPR000504	RNA recognition mc
	222	320	4.2E-29	T	21-09-2022	IPR012677	Nucleotide-binding
	321	431	1.9E-30	T	21-09-2022	IPR012677	Nucleotide-binding

What's the Pfam E-value distribution ?

1/ select Pfam hit lines

The screenshot shows the Galaxy France web interface. The top navigation bar includes the Galaxy France logo, a home icon, and links for Workflow, Visualize, Données partagées, Aide, Utilisateur, and a notification bell. A maintenance notice is displayed below the navigation bar. The left sidebar contains a 'Tools' section with a search bar and an 'Upload Data' button. Under 'Collection Operations', the 'GENERAL TEXT TOOLS' section is expanded, showing various text manipulation tools. The tool 'Select lines that match an expression' is highlighted with a red circle. The main content area shows the configuration for this tool, including a dropdown menu for the input data (1984: InterProScan on data 3 (tsv)), a dropdown for the matching method (Matching), and a text input field for the pattern (Pfam). There are also checkboxes for 'Keep header line' and 'Email notification', both currently set to 'No'. An 'Execute' button is visible at the bottom of the configuration area. A tip is provided: 'TIP: If your data is not TAB delimited, use Text Manipulation->Convert'. The 'Syntax' section at the bottom explains that the tool searches for lines containing or not containing a match to the given pattern.

Galaxy France Workflow Visualize Données partagées Aide Utilisateur

! UseGalaxy.fr will be undergoing maintenance from October 6th to 7th. Running jobs will be stopped. Thank you for your understanding.

Tools ☆ ☰

search tools ✕

Upload Data

Collection Operations

GENERAL TEXT TOOLS

Text Manipulation

Filter and Sort

- Filter sequences by ID from a tabular file
- Column arrange by header name
- Sub-sample sequences files e.g. to reduce coverage
- Filter data on any column using simple expressions
- Sort data in ascending or descending order
- Select lines that match an expression**
- Extract features from GFF data
- Filter GFF data by attribute using simple expressions
- Filter GFF data by feature count using simple expressions

Select lines that match an expression (Galaxy Version 1.0.3) ☆ ▾

Select lines from

1984: InterProScan on data 3 (tsv) ⬇

that

Matching ▾

the pattern

Pfam

here you can enter text or regular expression (for syntax check lower part of this frame)

Keep header line

No

i.e. the first line is kept independent of the regular expression

Email notification

No

Send an email notification when the job completes.

Execute









TIP: If your data is not TAB delimited, use *Text Manipulation->Convert*

Syntax



The select tool searches the data for lines containing or not containing a match to the given pattern. Regular Expression is introduced in this tool.


What's the Pfam E-value distribution ?


1/ select Pfam hit lines

Galaxy France [Workflow](#) [Visualize](#) [Données partagées](#) [Aide](#) [Utilisateur](#)        

! UseGalaxy.fr will be undergoing maintenance from October 6th to 7th. Running jobs will be stopped. Thank you for your understanding.

Tools  

search tools 

 Upload Data

Collection Operations

GENERAL TEXT TOOLS

Text Manipulation

Filter and Sort

- Filter sequences by ID** from a tabular file
- Column arrange** by header name
- Sub-sample sequences files** e.g. to reduce coverage
- Filter data** on any column using simple expressions
- Sort data** in ascending or descending order

FUN_008567-T1	30f0e5f7e973401301eef93092cc7760	122	Pfam	PF00076	RNA recognition motif. (a.k.a. RRM, RBD, or RNP domain)
FUN_003130-T1	bd4bab44ef23e2cf17f3e0d1c992e018	634	Pfam	PF00076	RNA recognition motif. (a.k.a. RRM, RBD, or RNP domain)
FUN_003130-T1	bd4bab44ef23e2cf17f3e0d1c992e018	634	Pfam	PF00076	RNA recognition motif. (a.k.a. RRM, RBD, or RNP domain)
FUN_003130-T1	bd4bab44ef23e2cf17f3e0d1c992e018	634	Pfam	PF00076	RNA recognition motif. (a.k.a. RRM, RBD, or RNP domain)
FUN_003130-T1	bd4bab44ef23e2cf17f3e0d1c992e018	634	Pfam	PF00076	RNA recognition motif. (a.k.a. RRM, RBD, or RNP domain)
FUN_003130-T1	bd4bab44ef23e2cf17f3e0d1c992e018	634	Pfam	PF00658	Poly-adenylate binding protein, unique domain
FUN_003678-T1	bd8e2066e65d046fe39d3ea94f2543ac	603	Pfam	PF02779	Transketolase, pyrimidine binding domain
FUN_003678-T1	bd8e2066e65d046fe39d3ea94f2543ac	603	Pfam	PF16870	2-oxoglutarate dehydrogenase C-terminal
FUN_003678-T1	bd8e2066e65d046fe39d3ea94f2543ac	603	Pfam	PF00676	Dehydrogenase E1 component
FUN_006750-T1	1c2b18a6e6d05afaf94e5d0b32a30128	554	Pfam	PF01853	MOZ/SAS family
FUN_006750-T1	1c2b18a6e6d05afaf94e5d0b32a30128	554	Pfam	PF17772	MYST family zinc finger domain
FUN_006750-T1	1c2b18a6e6d05afaf94e5d0b32a30128	554	Pfam	PF00628	PHD-finger
FUN_002211-T1	353fc52143176928bca800dc491d4e8d	347	Pfam	PF10516	SHNi-TPR
FUN_002293-T1	5d490c0e068c14797213b1e58c847038	829	Pfam	PF13193	AMP-binding enzyme C-terminal domain
FUN_002293-T1	5d490c0e068c14797213b1e58c847038	829	Pfam	PF16177	Acetyl-coenzyme A synthetase N-terminus
FUN_002293-T1	5d490c0e068c14797213b1e58c847038	829	Pfam	PF00501	AMP-binding enzyme
FUN_005342-T1	2931a1f5db4fe7f7cb61223895be49c6	177	Pfam	PF05255	Uncharacterised protein family (UPF0220)
FUN_007314-T1	014faa6d38986eada2186be73a052c1c	368	Pfam	PF00704	Glycosyl hydrolases family 18
FUN_007400-T1	883d2aec020c728e5f15a6c339b9e5d8	921	Pfam	PF03707	Bacterial signalling protein N terminal repeat
FUN_007400-T1	883d2aec020c728e5f15a6c339b9e5d8	921	Pfam	PF03707	Bacterial signalling protein N terminal repeat

What's the Pfam E-value distribution ?

1/ Compute and draw a histogram

Galaxy France Workflow Visualize Données partagées Aide Utilisateur

! UseGalaxy.fr will be undergoing maintenance from October 6th to 7th. Running jobs will be stopped. Thank you for your understanding.

Tools ☆ ☰

histogram ✕

Upload Data

Show Sections

Draw nucleotides distribution chart

EstimateLibraryComplexity assess sequence library complexity from read sequences

dada2: plotComplexity Plot sequence complexity profile

Circos: Link Density Track reduce links to a density plot

Histogram of a numeric column

SARTools edgeR Compare two or more biological conditions in a RNA-Seq framework with edgeR

SARTools DESeq2 Compare two or more biological conditions in a RNA-Seq framework with DESeq2

Histogram of a numeric column

MSI Qualitycontrol mass spectrometry imaging QC

Histogram of a numeric column (Galaxy Version 1.0.4) ☆ ▾

Dataset

1990: Select on data 1984

Dataset missing? See TIP below

Numerical column for x axis

Column: 9

Number of breaks (bars)

0

Plot title

Pfam E-value distribution

Label for x axis

E-value

Include smoothed density

No

Plot as frequency (counts)

No

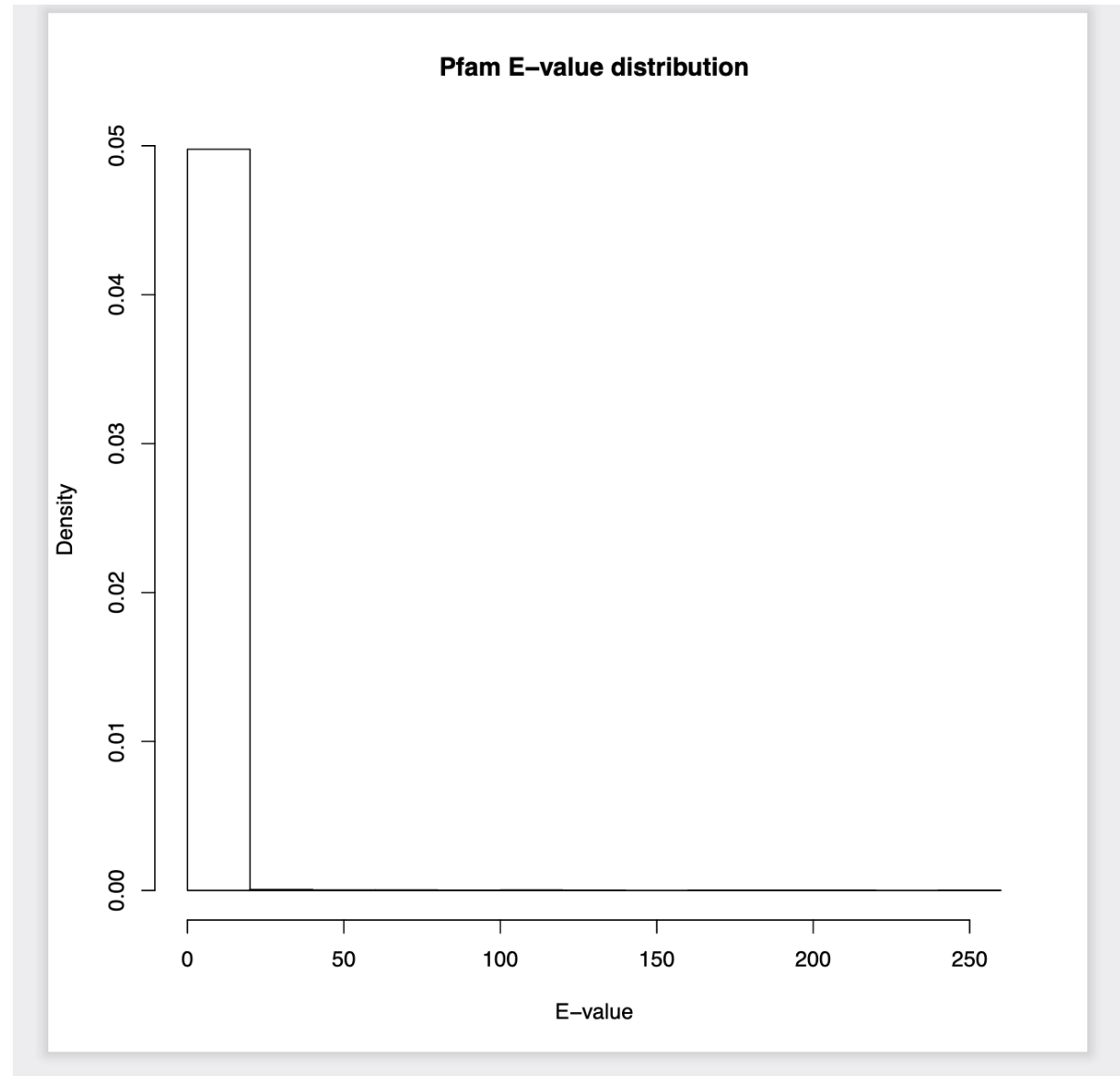
Email notification

No

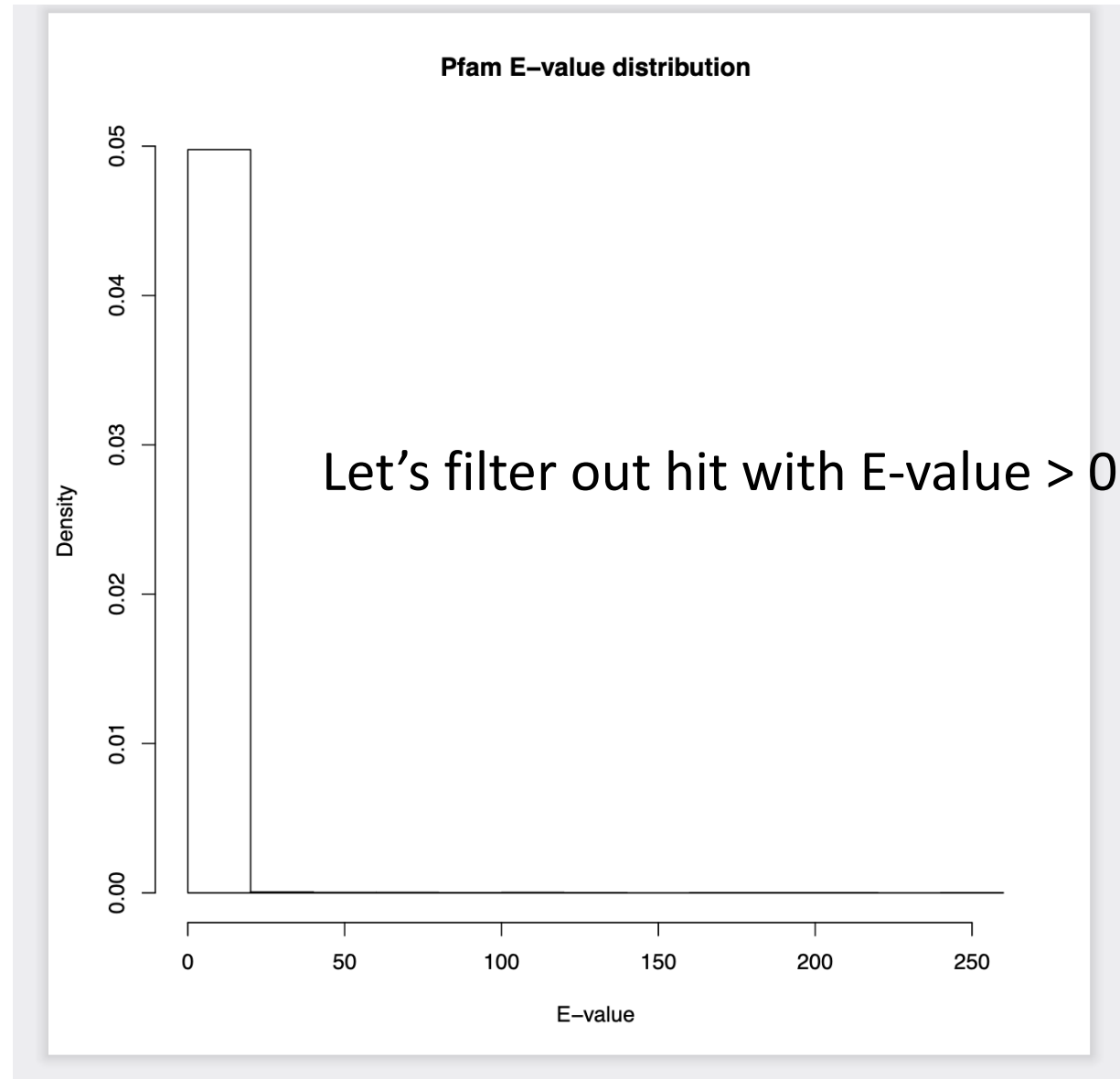
Send an email notification when the job completes.

Execute

What's the Pfam E-value distribution ?



What's the Pfam E-value distribution ?



Let's filter out hit with E-value > 0.001 !

Galaxy France Workflow Visualize Données partagées Aide Utilisateur

! UseGalaxy.fr will be undergoing maintenance from October 6th to 7th. Running jobs will be stopped. Thank you for your understanding.

Tools ☆ ☰

search tools ✕

Upload Data

Get Data

Send Data

Collection Operations

GENERAL TEXT TOOLS

Text Manipulation

Filter and Sort

- Filter sequences by ID from a tabular file
- Column arrange by header name
- Sub-sample sequences files e.g. to reduce coverage
- Filter data on any column using simple expressions**
- Sort data in ascending or descending order

Filter data on any column using simple expressions (Galaxy Version 1.1.1) ☆ ▾

Filter

1990: Select on data 1984 ⬇ ⬆ 📁

Dataset missing? See TIP below.

With following condition

c9<0.001

Double equal signs, ==, must be used as shown above. To filter for an arbitrary string, use the Select tool.

Number of header lines to skip

0 ⬆

Email notification

No

Send an email notification when the job completes.

Execute

⚠ Double equal signs, ==, must be used as "equal to" (e.g., c1 == 'chr22')

ℹ **TIP:** Attempting to apply a filtering condition may throw exceptions if the data type (e.g., string, integer) in every line of the columns being filtered is not appropriate for the condition (e.g., attempting certain numerical calculations on strings). If an exception is thrown when applying the condition to a line, that line is skipped as invalid for the filter condition. The number of invalid skipped lines is documented in the resulting history

What's the Pfam E-value distribution ?

Galaxy France Workflow Visualize Données partagées Aide Utilisateur

! UseGalaxy.fr will be undergoing maintenance from October 6th to 7th. Running jobs will be stopped. Thank you for your understanding.

Tools ☆ ☰

histogram ✕

Upload Data

Show Sections

Inner Distance calculate the inner distance (or insert size) between two paired RNA reads

Draw nucleotides distribution chart

EstimateLibraryComplexity assess sequence library complexity from read sequences

dada2: plotComplexity Plot sequence complexity profile

Circos: Link Density Track reduce links to a density plot

Histogram of a numeric column

SARTools edgeR Compare two or more biological conditions in a RNA-Seq framework with edgeR

SARTools DESeq2 Compare two or more biological conditions in a RNA-Seq framework with DESeq2

Histogram of a numeric column

MSI Qualitycontrol mass spectrometry

Histogram of a numeric column (Galaxy Version 1.0.4) ☆ ▾

Dataset

📄 📂 📁 1992: Filter on data 1990 ⬇ ⬆ 📁

Dataset missing? See TIP below

Numerical column for x axis

Column: 9 ▾

Number of breaks (bars)

0 ⬆

Plot title

Pfam filtered hits E-value distribution

Label for x axis

E-value

Include smoothed density

Yes

Plot as frequency (counts)

No

Email notification

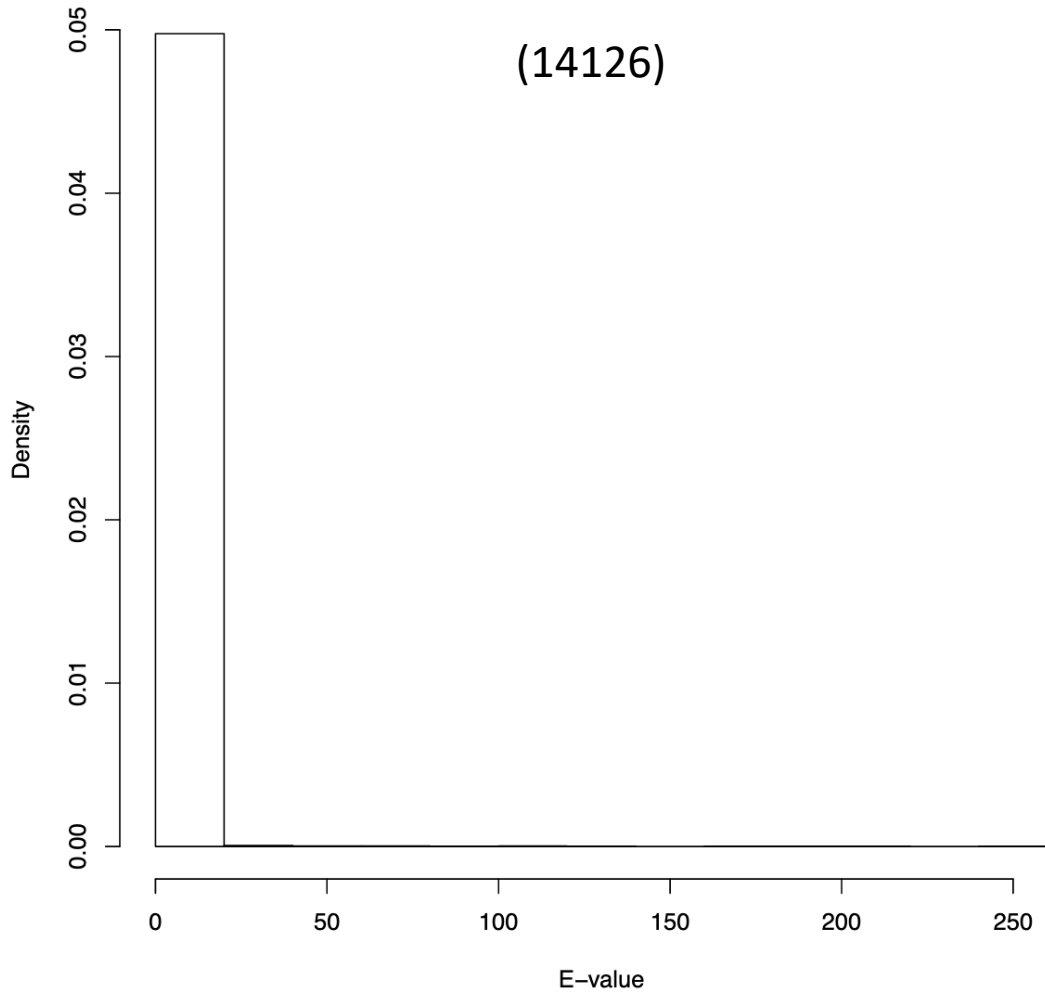
No

Send an email notification when the job completes.

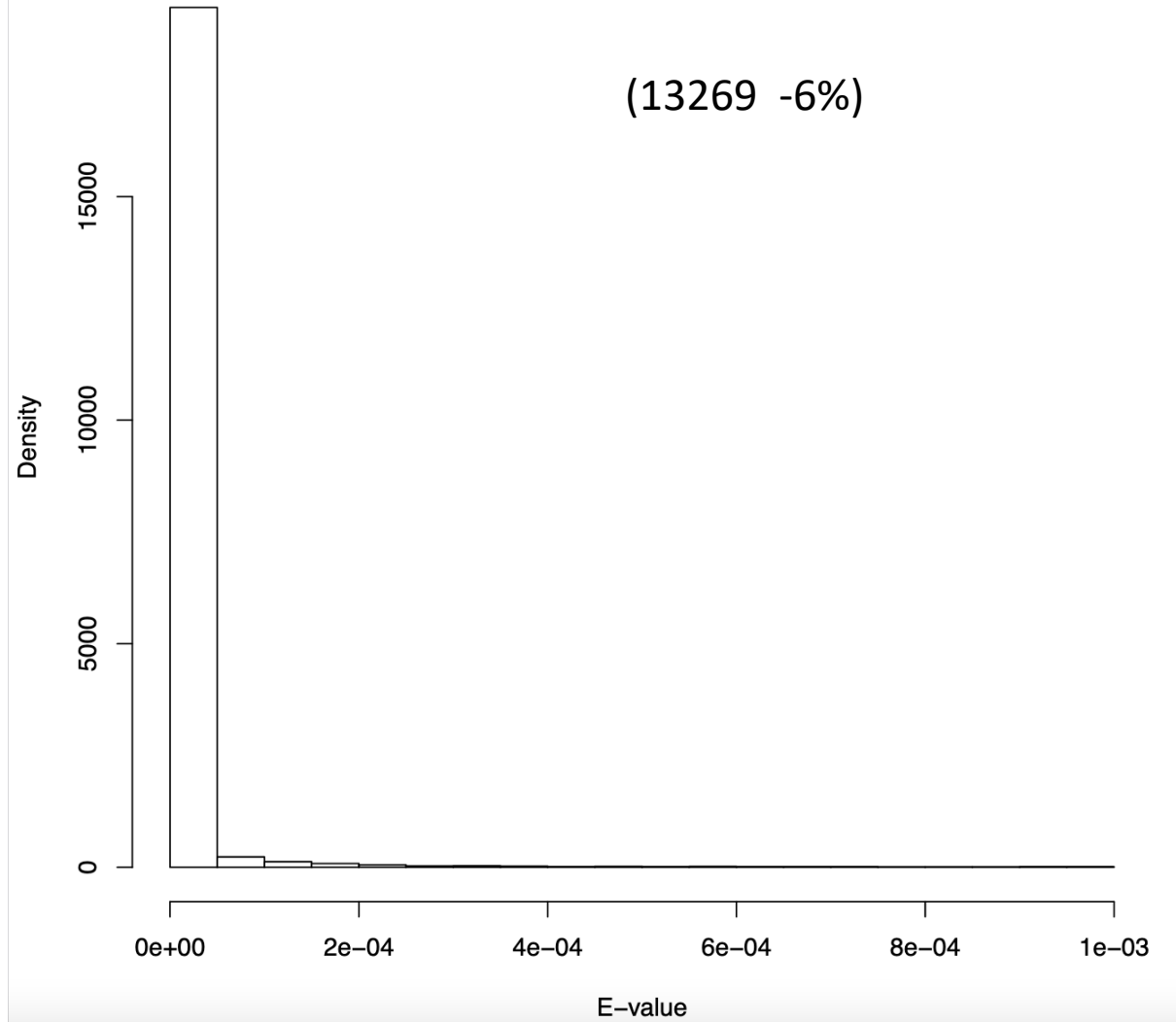
Execute

What's the Pfam E-value distribution ?

Pfam E-value distribution



Pfam filtered hits E-value distribution



TIPS : the Split by group tool :

Split hits into one file per database

Galaxy France Workflow Visualize Données partagées Aide Utilisateur Using 11%

! UseGalaxy.fr will be undergoing maintenance from October 6th to 7th. Running jobs will be stopped. Thank you for your understanding.

Tools

split Upload Data Show Sections

(split_libraries_fastq)

LAST-split finds "split alignments" (typically for DNA) or "spliced alignments" (typically for RNA).

Split libraries according to barcodes specified in mapping file (split_libraries)

bcftools norm Left-align and normalize indels; check if REF alleles match the reference; split multiallelic sites into multiple rows; recover multiallelics from multiple rows

Split by group

Run split_libraries_fastq on multiple files (multiple_split_libraries_fastq)

Split.abund Separate sequences into rare and abundant groups

Split file to dataset collection

Split.groups Generates a fasta file for each group

Barcode Splitter

bcftools mpileup Generate VCF or BCF

Split by group (Galaxy Version 0.6)

File to split
1992: Filter on data 1990

on column
Column: 1

Include header in splits?
 No
Include the first line (the assumed header line) to all split files.

Email notification
 No
Send an email notification when the job completes.

Execute

Synopsis

Given a single input dataset this tool splits the file on unique values from a specified column.

Description

This tool splits a file into a collection based on unique values of a specific column. It performs a grouping operation with every group saved as a separate collection element. You have the option to include the header (first line) to all splits. If you have a header and don't want keep it, please remove it before you use this tool. For example with the "Remove beginning of a file" tool.

Example

Splittng this file on column 1:

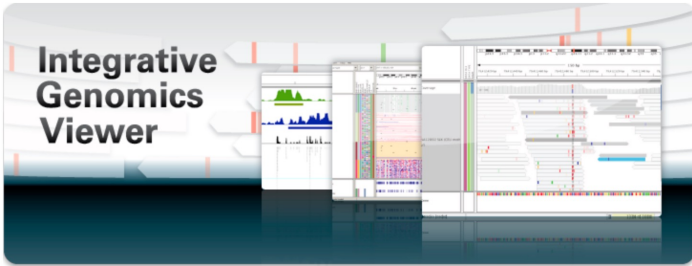
History

[Back to MucorProtSet](#)

Split by group collection
a list with 14 items

CDD		
Coils		
Gene3D		
Hamap		
PANTHER		
PIRSF		
PRINTS		
Pfam		
ProSitePatterns		
ProSiteProfiles		
SFLD		
SMART		
SUPERFAMILY		

InterProScan GFF3 output can be imported into IGV



<https://software.broadinstitute.org/software/igv/>

History ↻ + ▢ ⚙

Rechercher des données ? ✕

MucorProtSet

43 shown, 73 deleted, 1881 hidden

11.45 GB ☑ 🗉 💬

1986: InterProScan on data 3 (gff3) 👁 ✎ ✕

~180,000 lines
format: **gff3**, génome de référence: ?

27/09/2022 07:04:54:135 Welcome to InterProScan-5.55-88.0
27/09/2022 07:04:54:136 Running InterProScan v5 in STANDALONE mode... on Linux
27/09/2022 07:05:00:226 RunID: cpu-node-38.ifb.local_20220927_07045996
27/09/2022 07:05:08:240 Loading file


/S/n

🗉 🔗 🔍 🔄 📄 ? 🗉 💬

display with IGV local

The screenshot shows the IGV interface with a genomic track. The top bar displays the current view: "Galaxy3-[Galaxy13-[Fun... FUN_000005-T1 FUN_000005-T1:1-131] Go". Below the bar is a scale from 0 to 120 bp. The main track shows a sequence: "KLHMRHPSARNDMRTHINCLYKDRBHEN SQSPMTLLLDICWPHLWICWYEQIPFLBQLCRYBIBPSLYCFCEABSPHTPTERCRCABSNHLLTCLLDLDLQIPABABICDHWCCTBAWYFCQFRPHYBMTLTSF". Below the sequence are several tracks of protein annotations, each represented by a red bar with arrows indicating the direction. The annotations include: "FUN_000005-T1", "G3DSA:3.40.50.1820", "PTHR43248:SF12", "PTHR43248", "SSF53474", "PF00561", "PR00412", and "PR00111". A pop-up window is open over the PR00412 annotation, showing details: "Type: protein_match", "date: 27-09-2022", "Target: FUN_000005-T1 39 57", "Ontology_term: \"GO:0003824\"", "ID: match\$25232_39_57", "signature_desc: Epoxide hydrolase signature", "Name: PR00412", "status: T", "Dbxref: \"InterPro:IPR000639\", \"MetaCyc:PWY-1822\", \"MetaCyc:PWY-321\", \"MetaCyc:PWY-5292\", \"MetaCyc:PWY-5319\", \"Met". The bottom status bar shows "2 tracks", "FUN_000005-T1:108", and "851M of 1 944M".

InterProScan JSON output can be imported into IGV

 Classification of protein families 🔍 ☰


[Home](#) | [Search](#) | [Browse](#) | **[Results](#)** | [Release notes](#) | [Download](#) | [Help](#) | [About](#)



[🏠](#) / [Result](#) / [InterProScan](#)



Your InterProScan Search Results ⁱ




Your InterProScan search results are shown below. Searches may take varying times to complete. You can navigate to other pages and once the search is finished, you will receive a notification. The results will be available for 7 days.

Alternatively, you can import the results of an InterProScan run (in JSON format) into this page in order to view your search results interactively.

[Submit a new search](#) 

Import:  

 [Clear All](#) 

 RESULTS	 NAME	 CREATED	STATUS	ACTION
No data available				

[Previous](#) **1** [Next](#)

InterProScan JSON output can be imported into IGV

The screenshot shows the InterProScan web interface. At the top, the logo 'InterPro' is visible, along with the text 'Classification of protein families'. A search bar and a menu icon are in the top right. The main content area is partially obscured by a modal dialog box titled 'InterProScan File'. The dialog box contains the following text:

Loading file: `test.json`

You are about to load the analysis of **3 sequences** with InterProScan version 5.55-88.0

Mismatched Version

InterProScan version: `5.55-88.0`.

Some links might not work as the results are from a previous release of InterPro `88.0` and some of the data might have been deleted or changed in the current version `90.0`.

An 'OK' button is located at the bottom right of the dialog box. In the background, the main page shows a search bar, a 'Submit a new search' button, and a table with columns for 'RESULTS', 'NAME', 'CREATED', 'STATUS', and 'ACTION'. The table currently displays 'No data available' and a page number '1' between 'Previous' and 'Next' navigation links.

InterProScan JSON output can be imported into IGV

🏠 / Result / InterProScan

Your InterProScan Search Results ⁱ

Your InterProScan search results are shown below. Searches may take varying times to complete. You can navigate to other pages and once the search is finished, you will receive a notification. The results will be available for 7 days.







Alternatively, you can import the results of an InterProScan run (in JSON format) into this page in order to view your search results interactively.

Submit a new search 

Import:  

1 - 3 of 3 results

 Clear All ▾

RESULTS	NAME	CREATED	STATUS	ACTION
imported_file-test.json-3	FUN_003182-T1	just now		
imported_file-test.json-2	FUN_001534-T1	just now		
imported_file-test.json-1	FUN_011528-T1	just now		

Previous 1 Next

InterProScan JSON output can be imported into IGV

The screenshot displays the InterPro web interface. At the top, the navigation bar includes the InterPro logo, the title "Classification of protein families", and search and menu icons. Below the navigation bar, the "Results" tab is active, showing details for a protein entry: Title (FUN_001534-T1), Job ID (imported_file-test.json-2), Length (249 amino acids), Actions (delete and refresh), and Status (imported file).

The "Protein family membership" section indicates "None predicted".


The "Entry matches to this protein" section features a sequence alignment viewer with a scale from 1 to 249. It includes zoom controls and buttons for "Options" and "Export".

Below the alignment viewer, two sections are visible: "Homologous Superfamily" and "Predictions".

Homologous Superfamily: Shows three matches with colored bars indicating alignment regions. The matches are: IPR036859: CAP-Gly_dom_s (blue bar), G3DSA:2.30.30.190: (green bar), and SSF74924: Cap-Gly domain (light green bar).

Predictions: Shows three matches with colored bars indicating alignment regions. The matches are: mobidb-lite (1) (teal bar), mobidb-lite (2) (pink bar), and mobidb-lite (3) (olive bar).

InterProScan JSON output can be imported into IGV



InterPro Classification of protein families

Home ▶ Search ▶ **Browse** ▶ Results Release notes Download ▶ Help ▶ About

🏠 / Browse / By Entry / InterPro / IPR036859 / Overview



CAP Gly-rich domain superfamily[★]

IPR036859

InterPro entry ⓘ



Short name: *CAP-Gly_dom_sf*

Overview

Proteins	28k
Taxonomy	8k
Proteomes	2k
Structures	45
AlphaFold	17k
Pathways	65
Genome3D	1k

Overlapping entries ⓘ

D [CAP Gly-rich domain](#) (IPR000938)

Description

Cytoskeleton-associated proteins (CAPs) are involved in the organisation of microtubules and transportation of vesicles and organelles along the cytoskeletal network. A conserved glycine-rich domain, CAP-Gly, has been identified in a number of CAPs, including CLIP-170 and dynactins. The crystal structure of the *Caenorhabditis elegans* F53F4.3 protein CAP-Gly domain has been solved. The domain contains three β -strands. The most conserved sequence, GKNDG, is located in two consecutive sharp turns on the surface, forming the entrance to a groove ^[1].

Functional annotation pipelines



OPINION ARTICLE

Ten steps to get started in Genome Assembly and Annotation

[version 1; peer review: 2 approved]

Victoria Dominguez Del Angel ¹, Erik Hjerde ², Lieven Sterck ^{3,4},
Salvadors Capella-Gutierrez ^{5,6}, Cederic Notredame^{7,8},
Olga Vinnere Pettersson⁹, Joelle Amselem ¹⁰, Laurent Bouri ¹,
Stephanie Bocs ¹¹⁻¹³, Christophe Klopp ¹⁴, Jean-Francois Gibrat ^{1,15},
Anna Vlasova ⁸, Brane L. Leskosek¹⁶, Lucile Soler¹⁷, Mahesh Binzer-Panchal ¹⁷,
Henrik Lantz ¹⁷

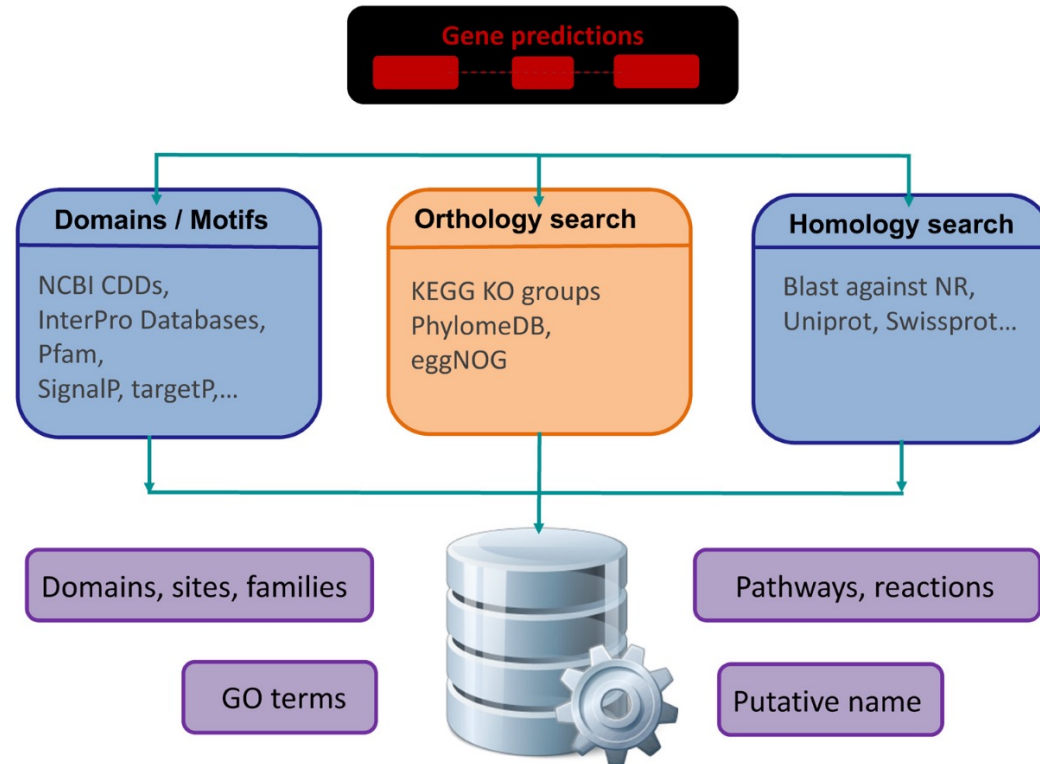


Figure 4. Functional Annotation Pipelines. This schema is showing a typical functional annotation pipeline, in which functional roles are assigned to coding sequences (CDSs) inferred in the gene prediction process. The process implements three parallel routes for the definition of functions. The first refers to proteins domains and motifs, the second for orthology search and finally the third is applied to homology search. At the end, the output from the three different sources is put together for more valuable predictions.

Orthology-based functional annotation with eggNOG

Fast Genome-Wide Functional Annotation through Orthology Assignment by eggNOG-Mapper

Jaime Huerta-Cepas,^{†,1} Kristoffer Forslund,^{†,1} Luis Pedro Coelho,¹ Damian Szklarczyk,^{2,3} Lars Juhl Jensen,⁴ Christian von Mering,^{2,3} and Peer Bork^{*,1,5,6,7}

¹Structural and Computational Biology Unit, European Molecular Biology Laboratory, Heidelberg, Germany

²Institute of Molecular Life Sciences, University of Zurich, Zurich, Switzerland

³Bioinformatics/Systems Biology Group, Swiss Institute of Bioinformatics (SIB), Zurich, Switzerland

⁴The Novo Nordisk Foundation Center for Protein Research, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark

⁵Germany Molecular Medicine Partnership Unit (MMPU), University Hospital Heidelberg and European Molecular Biology Laboratory, Heidelberg, Germany

⁶Max Delbrück Centre for Molecular Medicine, Berlin, Germany

⁷Department of Bioinformatics, Biocenter University of Würzburg, Würzburg, Germany

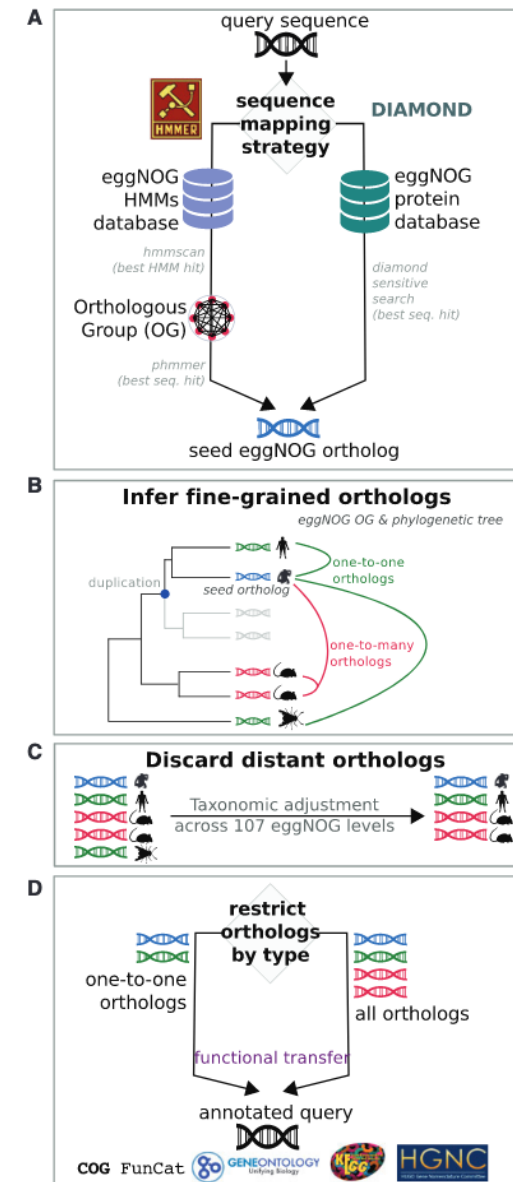
[†]These authors contributed equally to this work.

*Corresponding author: E-mail: bork@embl.de.

Associate editor: Hongzhi Kong

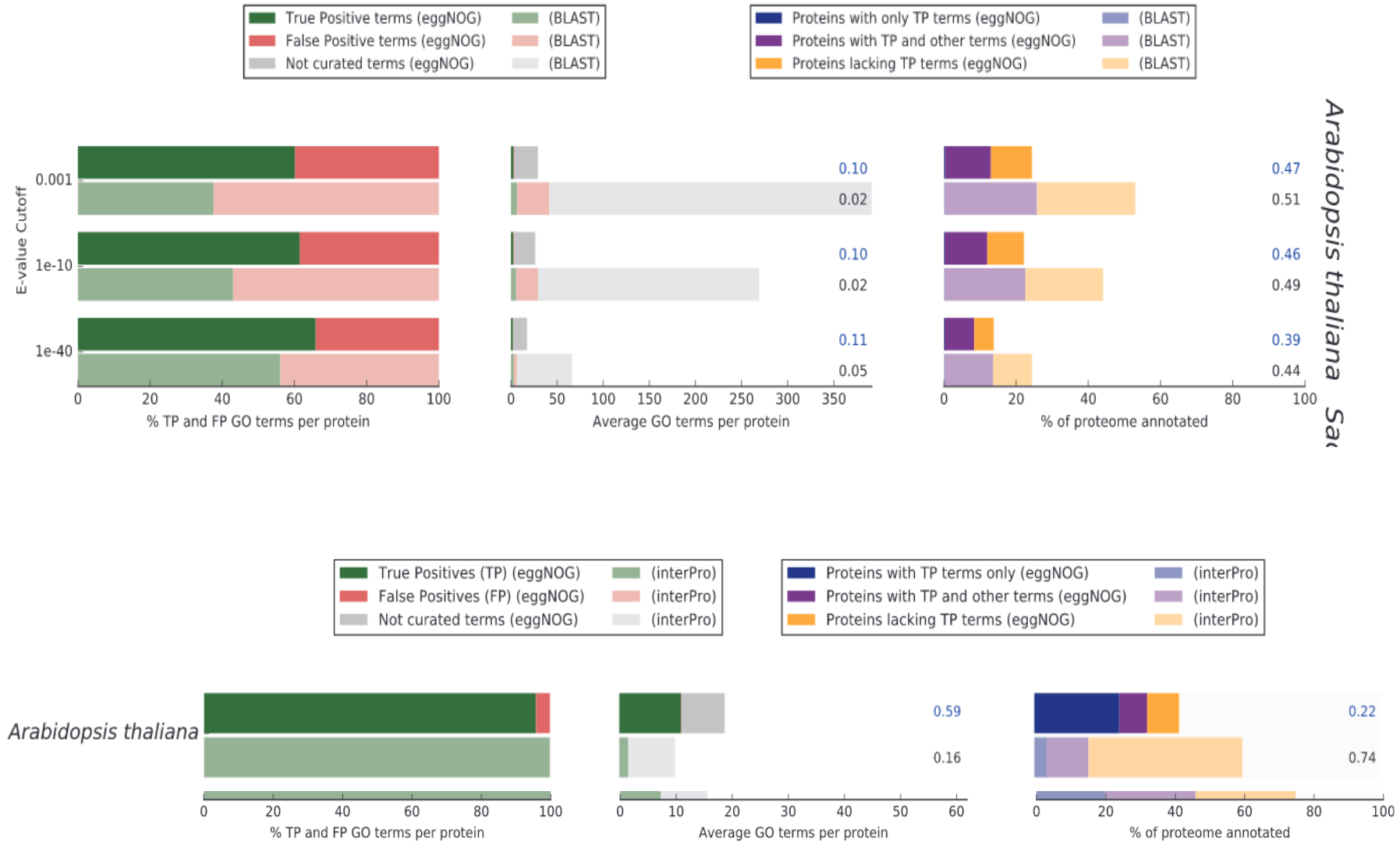
Abstract

Orthology assignment is ideally suited for functional inference. However, because predicting orthology is computationally intensive at large scale, and most pipelines are relatively inaccessible (e.g., new assignments only available through database updates), less precise homology-based functional transfer is still the default for (meta-)genome annotation. We, therefore, developed eggNOG-mapper, a tool for functional annotation of large sets of sequences based on fast orthology assignments using precomputed clusters and phylogenies from the eggNOG database. To validate our method, we benchmarked Gene Ontology (GO) predictions against two widely used homology-based approaches: BLAST and InterProScan. Orthology filters applied to BLAST results reduced the rate of false positive assignments by 11%, and increased the ratio of experimentally validated terms recovered over all terms assigned per protein by 15%. Compared with InterProScan, eggNOG-mapper achieved similar proteome coverage and precision while predicting, on average, 41 more terms per protein and increasing the rate of experimentally validated terms recovered over total term assignments per protein by 35%. EggNOG-mapper predictions scored within the top-5 methods in the three GO categories using the CAFA2 NK-partial benchmark. Finally, we evaluated eggNOG-mapper for functional annotation of metagenomics data, yielding better performance than InterProScan. eggNOG-mapper runs $\sim 15\times$ faster than BLAST and at least $2.5\times$ faster than InterProScan. The tool is available standalone and as an online service at <http://eggno-mapper.embl.de>.



Orthology-based functional annotation with eggNOG

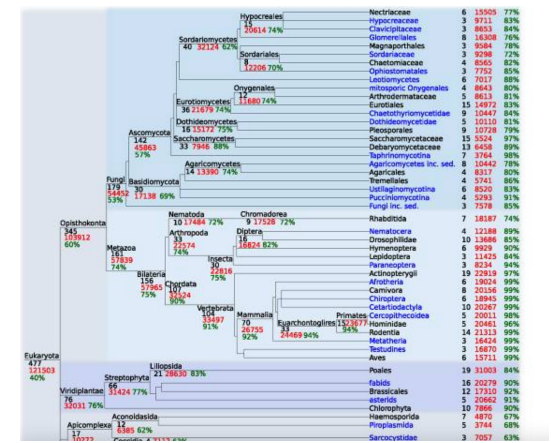
eggNOG is **more specific** but **less sensitive** than similarity (blast) or domains-based (InterProScan) methods



Let's have a look at eggNOG-mapper results

October 6th to 7th. Running jobs will be stopped. Thank you for your understanding.

FUN_000115-T1	36080.S2J093	1.52e-51	187.0	KOG1039@1 root,KOG1039@2759 Eukaryota,3958Y@33154 Opisthokonta,3Q2CW@4751 Fungi,1GXEC@112252 Fungi incertae sedis	4751 Fungi	S	protein ubiquitination
FUN_000116-T1	36080.S2JKD5	4.51e-41	163.0	KOG2992@1 root,KOG2992@2759 Eukaryota,38DH8@33154 Opisthokonta,3NUEP@4751 Fungi,1GTTM@112252 Fungi incertae sedis	4751 Fungi	Y	SRP40, C-terminal domain
FUN_000117-T1	36080.S2J4G6	0.0	1245.0	KOG3836@1 root,KOG3836@2759 Eukaryota,39JZT@33154 Opisthokonta,3P07N@4751 Fungi,1GVA9@112252 Fungi incertae sedis	4751 Fungi	K	ig-like, plexins, transcription factors
FUN_000118-T1	36080.S2JWH6	5.42e-25	115.0	2CWJA@1 root,2RUPX@2759 Eukaryota,39PS2@33154 Opisthokonta,3Q60Z@4751 Fungi,1GV0X@112252 Fungi incertae sedis	36080.S2JWH6	-	-
FUN_000119-T1	36080.S2K4Z5	7.46e-185	556.0	293MF@1 root,2RAI8@2759 Eukaryota,3AWX@33154 Opisthokonta,3Q1QD@4751 Fungi,1GVJ5@112252 Fungi incertae sedis	4751 Fungi	-	-
FUN_000120-T1	556268.OFAG_00919	1.17e-07	62.4	COG0790@1 root,COG0790@2 Bacteria,1MWA@1224 Proteobacteria,2VM3A@28216 Betaproteobacteria	1224 Proteobacteria	KLT	PFAM Sel1 domain protein repeat-containing protein
FUN_000121-T1	36080.S2JT78	6.53e-171	487.0	COG0564@1 root,KOG1919@2759 Eukaryota,38G5C@33154 Opisthokonta,3NVMQ@4751 Fungi,1GUC6@112252 Fungi incertae sedis	4751 Fungi	A	RNA pseudouridylate synthase
FUN_000122-T1	36080.S2IYX5	5.74e-196	556.0	KOG4733@1 root,KOG4733@2759 Eukaryota,38GX0@33154 Opisthokonta,3NWWZ@4751 Fungi,1GUJ5@112252 Fungi incertae sedis	4751 Fungi	A	RNA recognition motif
FUN_000123-T1	936053.I1CPJ9	3.37e-39	151.0	2DKYR@1 root,2S63V@2759 Eukaryota,3A7H5@33154 Opisthokonta,3P555@4751 Fungi	4751 Fungi	S	Spizellomyces punctatus DAOM BR117
FUN_000127-T1	36080.S2JU99	7.61e-139	413.0	KOG1819@1 root,KOG1819@2759 Eukaryota,39204@33154 Opisthokonta,3PHJW@4751 Fungi,1GVQF@112252 Fungi incertae sedis	4751 Fungi	S	negative regulation of epidermal growth factor-activated
FUN_000128-T1	36080.S2JUA4	1.79e-103	305.0	2CXIS@1 root,2RXWC@2759 Eukaryota,3A19Z@33154 Opisthokonta	33154 Opisthokonta	S	DUF2407 C-terminal domain
FUN_000129-T1	36080.S2I2Z8	9.73e-294	813.0	KOG1087@1 root,KOG1087@2759 Eukaryota,38DR4@33154 Opisthokonta,3NV9I@4751 Fungi,1GSCY@112252 Fungi incertae sedis	4751 Fungi	U	Adaptin C-terminal domain
FUN_000130-T1	36080.S2J5C2	4.29e-222	617.0	KOG1164@1 root,KOG1164@2759 Eukaryota,38BQ1@33154 Opisthokonta,3NUPD@4751 Fungi,1GV4X@112252 Fungi incertae sedis	4751 Fungi	T	Protein tyrosine kinase
FUN_000136-T1	36080.S2K6E0	5.88e-163	491.0	KOG4628@1 root,KOG4628@2759 Eukaryota,3AX63@33154 Opisthokonta,3Q20B@4751 Fungi,1GWRB@112252 Fungi incertae sedis	4751 Fungi	O	zinc-RING finger domain
FUN_000139-T1	36080.S2J1M7	1.29e-231	649.0	KOG3780@1 root,KOG3780@2759 Eukaryota,3AWX9@33154 Opisthokonta,3Q1PR@4751 Fungi,1GVFT@112252 Fungi incertae sedis	4751 Fungi	S	Arrestin (or S-antigen), C-terminal domain
FUN_000140-T1	36080.S2JXA7	8.04e-66	234.0	KOG4339@1 root,KOG4339@2759 Eukaryota,39S5X@33154 Opisthokonta,3NVE2@4751 Fungi,1GUWE@112252 Fungi incertae sedis	4751 Fungi	S	to Saccharomyces cerevisiae BNI4 (YNL233W)
FUN_000141-T1	36080.S2JLS1	1.94e-167	476.0	KOG0588@1 root,KOG0588@2759 Eukaryota,38BGN@33154 Opisthokonta,3NW2V@4751 Fungi,1GWAH@112252 Fungi incertae sedis	4751 Fungi	D	Serine/Threonine protein kinases, catalytic domain
FUN_000142-T1	36080.S2JKN5	7.38e-157	446.0	COG0445@1 root,KOG2667@2759 Eukaryota,39RPO@33154 Opisthokonta,3NUJ6@4751 Fungi,1GUF6@112252 Fungi incertae sedis	4751 Fungi	U	Endoplasmic reticulum vesicle transporter
FUN_000143-T1	936053.I1BJ13	4.18e-60	194.0	2D8F6@1 root,2T9ZJ@2759 Eukaryota,394TC@33154 Opisthokonta,3Q1W1@4751 Fungi,1GW90@112252 Fungi incertae sedis	4751 Fungi	-	-
FUN_000144-T1	36080.S2J4M1	1.23e-180	503.0	COG0090@1 root,KOG2309@2759 Eukaryota,38BCB@33154 Opisthokonta,3NU8K@4751 Fungi,1GSN6@112252 Fungi incertae sedis	4751 Fungi	J	Mortierella verticillata NRRL 6337 60S ribosomal protein
FUN_000145-T1	36080.S2JQW6	4.03e-281	775.0	COG2940@1 root,KOG4056@1 root,KOG2084@2759 Eukaryota,KOG4056@2759 Eukaryota,3A1HC@33154 Opisthokonta,3P2QG@4751 Fungi,1GRT3@112252 Fungi incertae sedis	4751 Fungi	BU	MAS20 protein import receptor
FUN_000146-T1	36080.S2JUA3	4.48e-224	622.0	COG3000@1 root,KOG0539@2759 Eukaryota,38D7Y@33154 Opisthokonta,3NV5W@4751 Fungi,1GS8X@112252 Fungi incertae sedis	4751 Fungi	I	Ceramide hydroxylase involved in the alpha-hydroxylation



annotations

This file provides final annotations of each query. Tab-delimited columns in the file are:

- query_name: query sequence name
- seed_eggNOG_ortholog: best protein match in eggNOG
- seed_ortholog_evalue: best protein match (e-value)
- seed_ortholog_score: best protein match (bit-score)
- predicted_taxonomic_group
- predicted_protein_name: Predicted protein name for query sequences
- GO_terms: Comma delimited list of predicted Gene Ontology terms
- EC_number
- KEGG_KO
- KEGG_Pathway: Comma delimited list of predicted KEGG pathways
- KEGG_Module
- KEGG_Reaction
- KEGG_rclass
- BRITE
- KEGG_TC
- CAZy
- BiGG_Reactions
- Annotation_tax_scope: The taxonomic scope used to annotate this query sequence
- Matching_OGs: Comma delimited list of matching eggNOG Orthologous Groups
- best_OG|evalue|score: Best matching Orthologous Groups (deprecated, use smallest from eggnog OGs)
- COG_functional_categories: COG functional category inferred from best matching OG
- eggNOG_free_text_description

INFORMATION STORAGE AND PROCESSING

- [J] Translation, ribosomal structure and biogenesis
- [A] RNA processing and modification
- [K] Transcription
- [L] Replication, recombination and repair
- [B] Chromatin structure and dynamics

CELLULAR PROCESSES AND SIGNALING

- [D] Cell cycle control, cell division, chromosome partitioning
- [Y] Nuclear structure
- [V] Defense mechanisms
- [T] Signal transduction mechanisms
- [M] Cell wall/membrane/envelope biogenesis
- [N] Cell motility
- [Z] Cytoskeleton
- [W] Extracellular structures
- [U] Intracellular trafficking, secretion, and vesicular transport
- [O] Posttranslational modification, protein turnover, chaperones

METABOLISM

- [C] Energy production and conversion
- [G] Carbohydrate transport and metabolism
- [E] Amino acid transport and metabolism
- [F] Nucleotide transport and metabolism
- [H] Coenzyme transport and metabolism
- [I] Lipid transport and metabolism
- [P] Inorganic ion transport and metabolism
- [Q] Secondary metabolites biosynthesis, transport and catabolism

POORLY CHARACTERIZED

- [R] General function prediction only
- [S] Function unknown