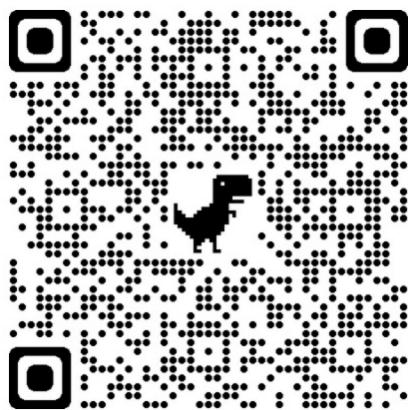




Lucile Soler PhD
Nima Rafati PhD

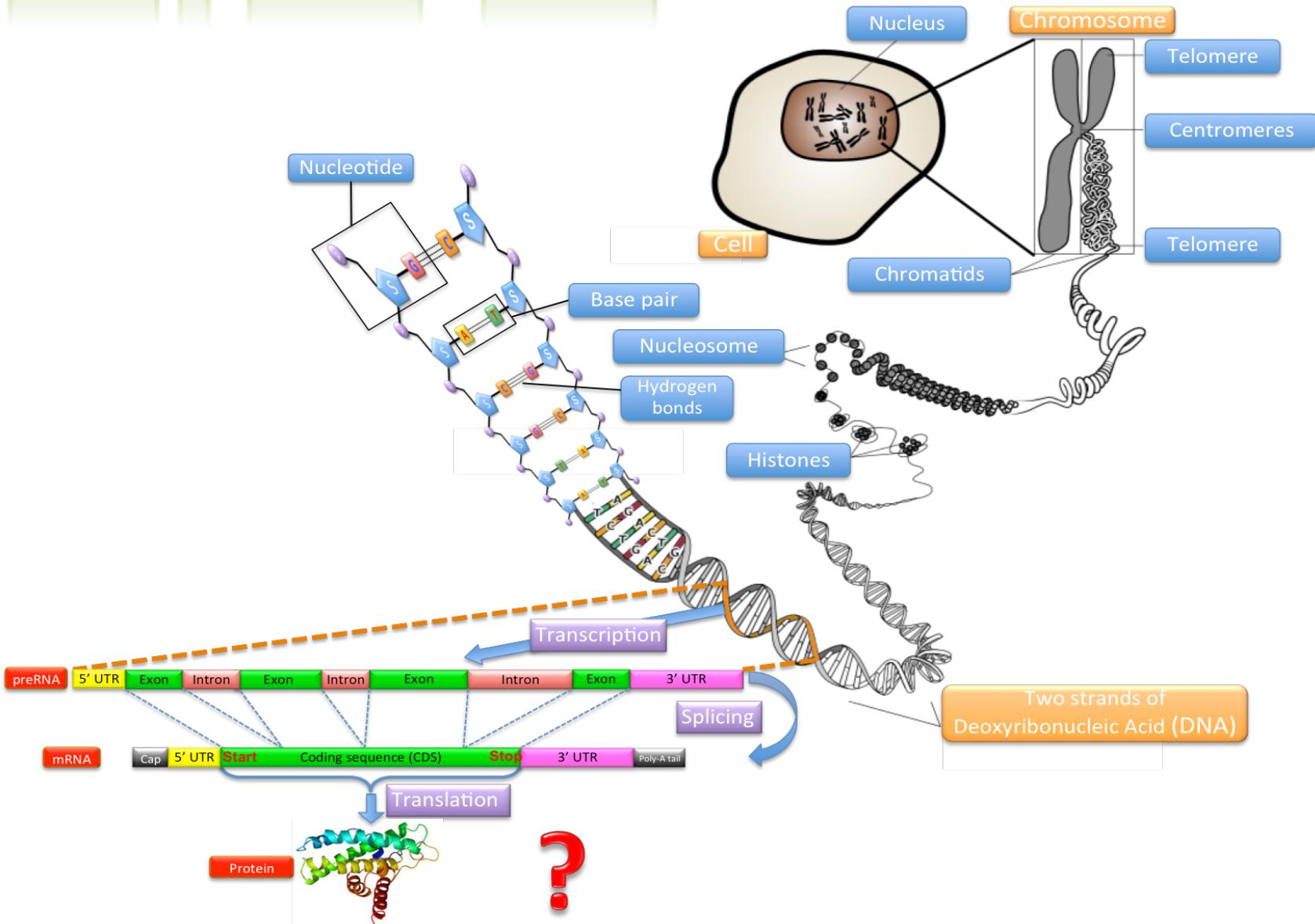
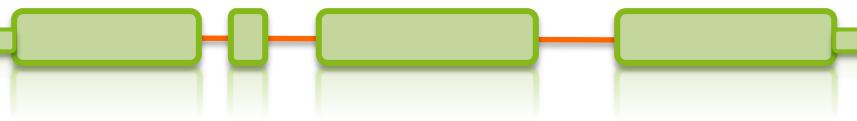
Functional annotation

https://nbisweden.github.io/workshop-genome_annotation_elixir/labs/functional_annotation



Genome assembly and annotation course
Norway May 2021

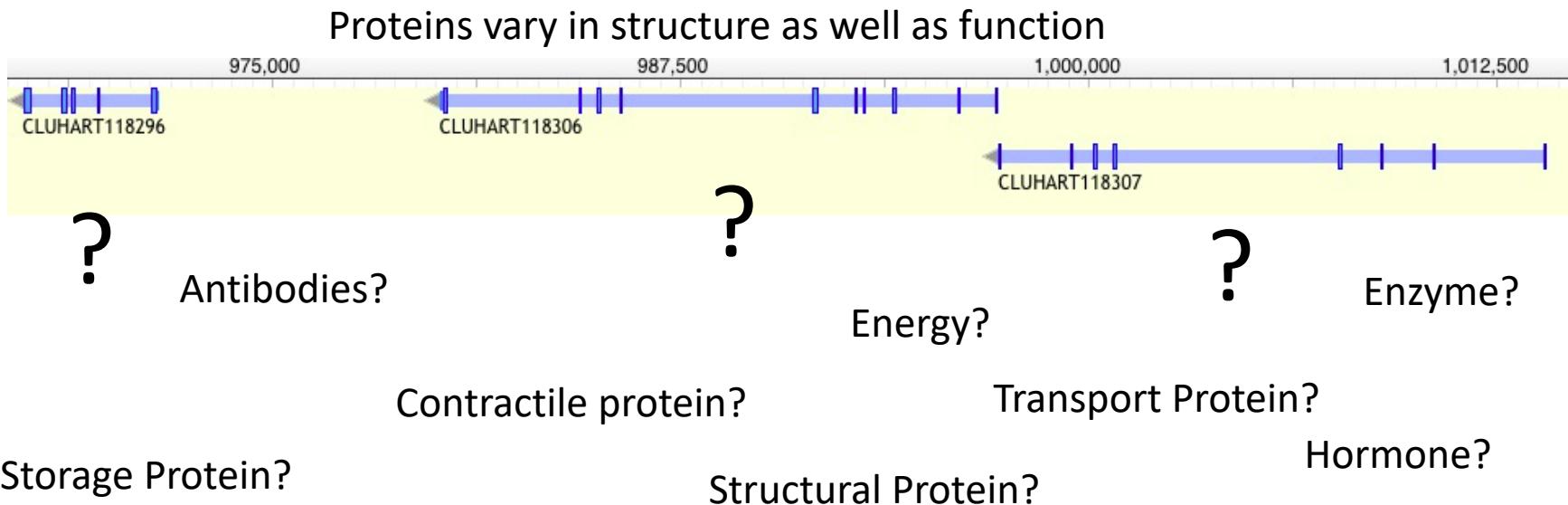
Overview



Functional annotation – WHY?

Understanding the function of gene product is key to understanding how a limited number of interacting gene products can generate life, from simple unicellular organisms to the incredibly complex multi-cellular *Homo sapiens*.

Rison,S.C., Hodgman,T.C. and Thornton,J.M. (2000) Comparison of functional annotation schemes for genomes. *Funct. Integr. Genomics*, 1, 56–69.



Functional annotation – HOW?

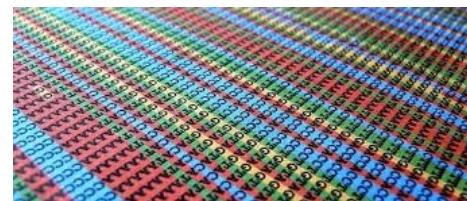


- Experimentally
=> Mutants, knockout, etc.
Accurate

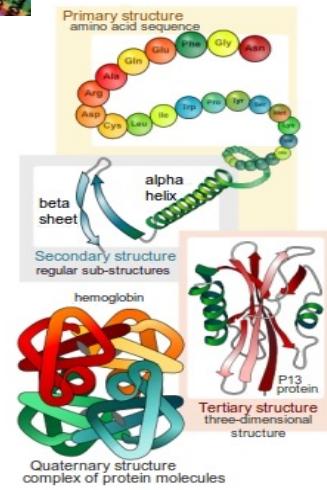


Mice homozygous for the diabetes 3J spontaneous mutation

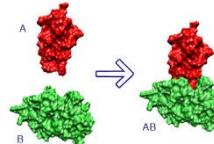
- Computationally
 - Sequence-based



- Structure based



- Protein-protein interaction data



limited accuracy

Methods - Sequence-based

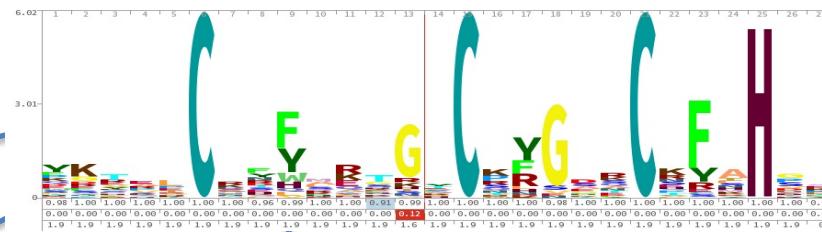
- Based on similarity
=>Best blast hit

Q GLMDTAFEHIKATGGLTTESNYPYKGEDATCNS-KJ
GLM+ AFE+IK +GG+TTES YPY+ + TC++ +
S GLMENAFEYIKHSGGITTESAYPYRAANGTCDAVR

- Based on Motif
⇒Proscan, MEME, QuasiMotifFinder

D-X-[KR]-P-{WYF}-X5

- Based on Profile (HMM or other statistical signature)



Whole sequence
e.g. Psi-BLAST*, PIRSF

domain
e.g. PFAM

Localization (e.g membrane, golgi, secreted)
e.g. SignalP, TMHMM

structural classification
e.g. SUPERFAMILY



- Based on evolutionary relationship (Orthology)
 - Clustering: KOG (Eukaryotic) / COG(Prokaryotic)
 - Based on synteny
 - ⇒ Whole genome alignment (lastZ)
 - (NBIS) Satsuma + kraken + custom script
 - Comparative Annotation Toolkit (CAT)
 - Based on phylogeny
 - ⇒ Quite complicated at large scale

- **Similarity to known structures.**

- Global structure-comparison
 - SCOP, CATH, and FSSP, are the most common methods used to classify protein structures (manual/automated)
- localized regions
 - might be relevant to function: clefts, pockets and surfaces
- active-site residues (catalytic clusters and ligand-binding sites)
 - active-site residues is often more conserved than the overall fold
⇒ PDBSiteScan

no single method is always successful

Functional annotation – HOW?

It is actually kind of complex...

- Multi-dimensional problem :
 - e.g. A protein can have a molecular function, a cellular role, and be part of a functional complex or pathway
- Molecular function can be illustrated by multiple descriptive levels
 - (e.g. '**enzyme**' category versus a more specific '**protease**' assignment).

Functional annotation – HOW?



It is actually kind of complex...

- Similarities (structural or in sequence)  function.
 - Similar sequence but different function (new domain => new combination => different function)
 - Different sequence may have same function (convergence) : Profiles helpful
 - Two proteins may have a similar fold but different functions
- Looks for conserved domains more reliable than whole sequence ?
 - How to go from conserved domains to assigning a function to your protein?

=> Importance is to gather as much information as possible

Sequence-based methods

- The most used (popular)
- Quick
- Easy to use
- Accurate (>70%)

Watson JD, Sanderson S, Ezersky A, Savchenko A, Edwards A, Orengo C, Joachimiak A, Laskowski RA, Thornton JM: Towards fully automated structure-based function prediction in structural genomics: a case study. *J Mol Biol.* 2007, 367: 1511-1522. 10.1016/j.jmb.2007.01.063.

- Many resources: even structural domains information
- Less computationally demanding

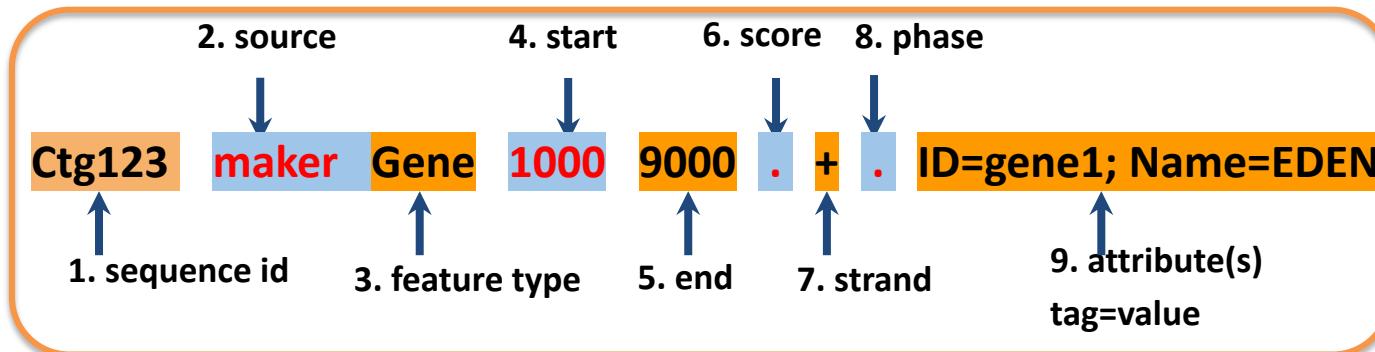
Functional annotation – HOW?



Get sequences

Functional annotation – HOW?

- Genome is in fasta format.
- Annotation is often in GFF-format. This format contains in general only coordinates, but sometimes it can include the sequence as well.



- You can use the GFF-file together with the genome-file to extract the gene sequences.
- The functional annotation tools want sequences in amino acid format, so when you extract the sequences you also need to convert the nucleotides to amino acids.

Functional annotation – HOW?



Get sequences



Search
similar
function

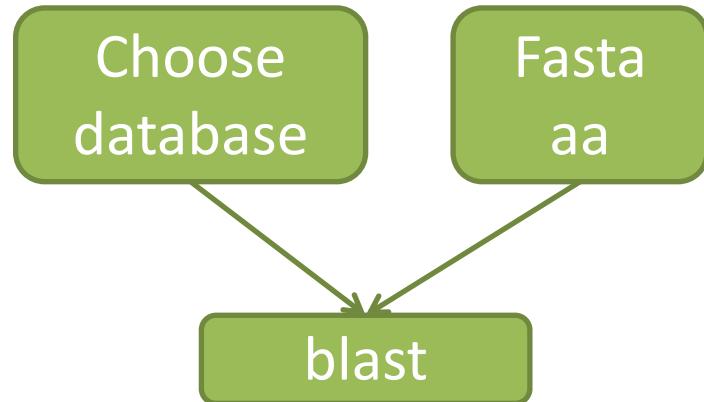
Blast-based
approach

Annotate the sequences functionally using Blast

Choose
database

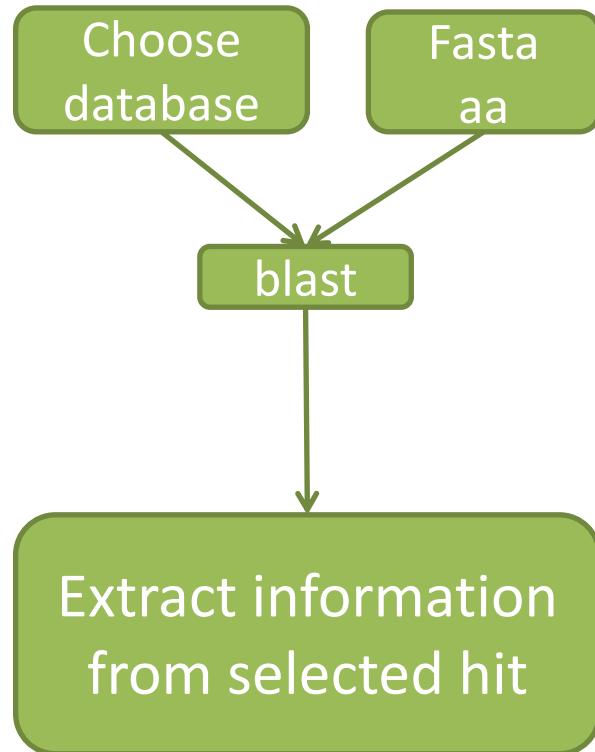
Uniprot	Swissprot
exhaustive	reliable

Annotate the sequences functionally using Blast



Minimum Threshold

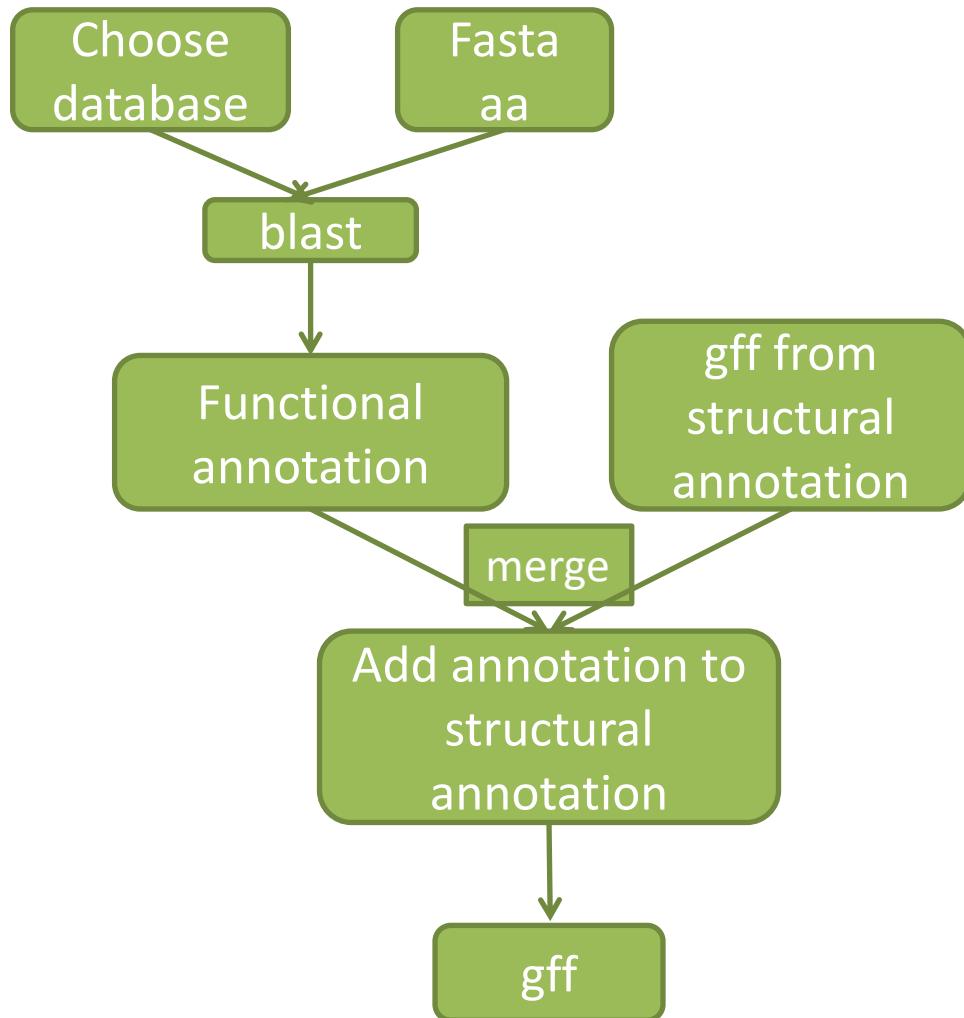
Annotate the sequences functionally using Blast



How to filter ?

- Minimum e-value
- Best blast hit
- You could prioritize by species

Annotate the sequences functionally using Blast



Blast-based approach

Strengths

- Fairly fast and easy
- Allow gene naming (e.g. plip)
- Overall function (e.g. Phosphatidylglycerophosphatase and protein-tyrosine phosphatase 1)

Limits

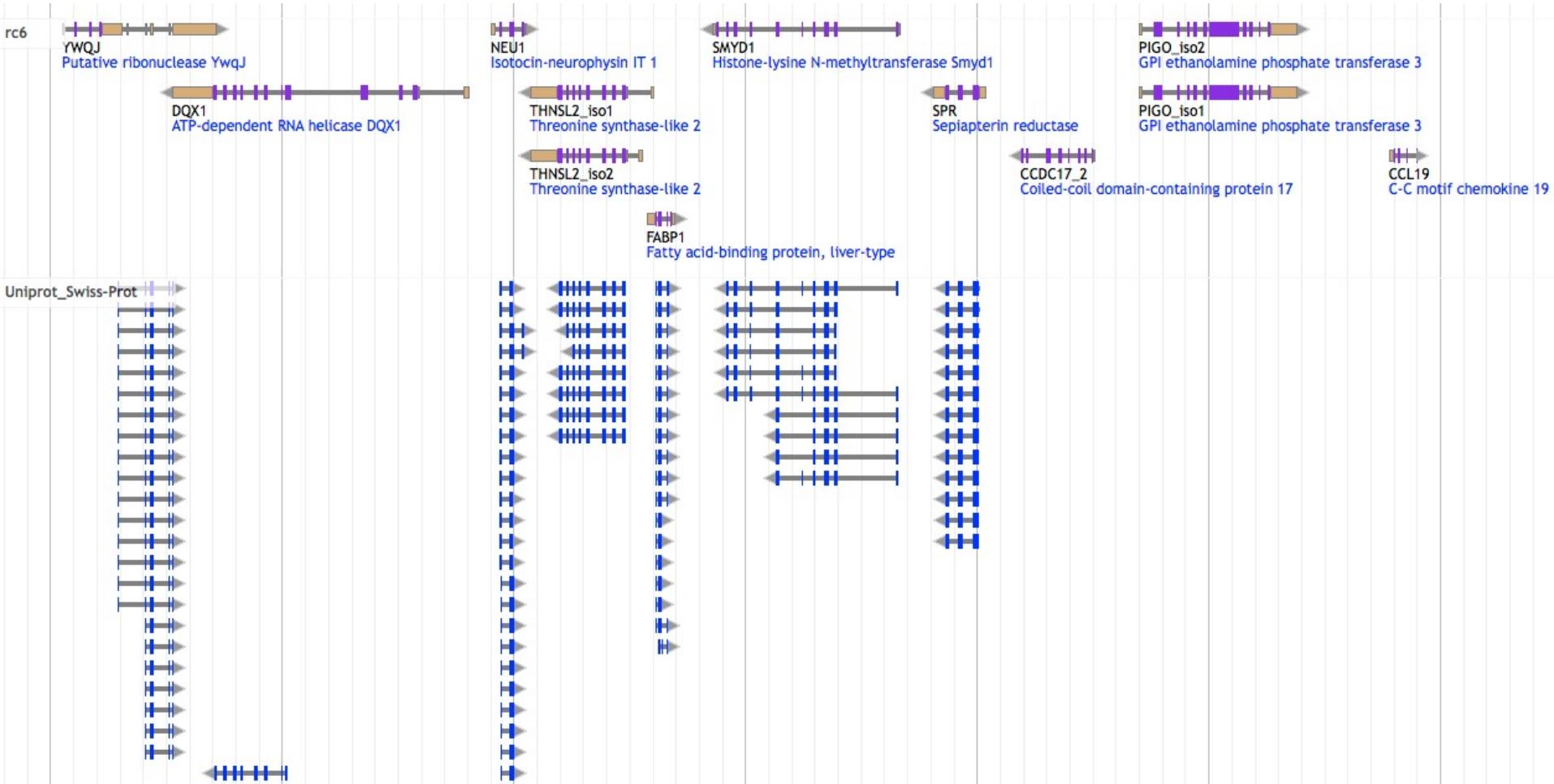
- Orthology not certain - best blast-hit does not equal orthologous!
- Bias due to well conserved domains
- Best Hit (use as template) is not necessarily the best annotated sequence to use => Could apply a prioritization rule (Human first, then mouse, etc).

Blast-based approach

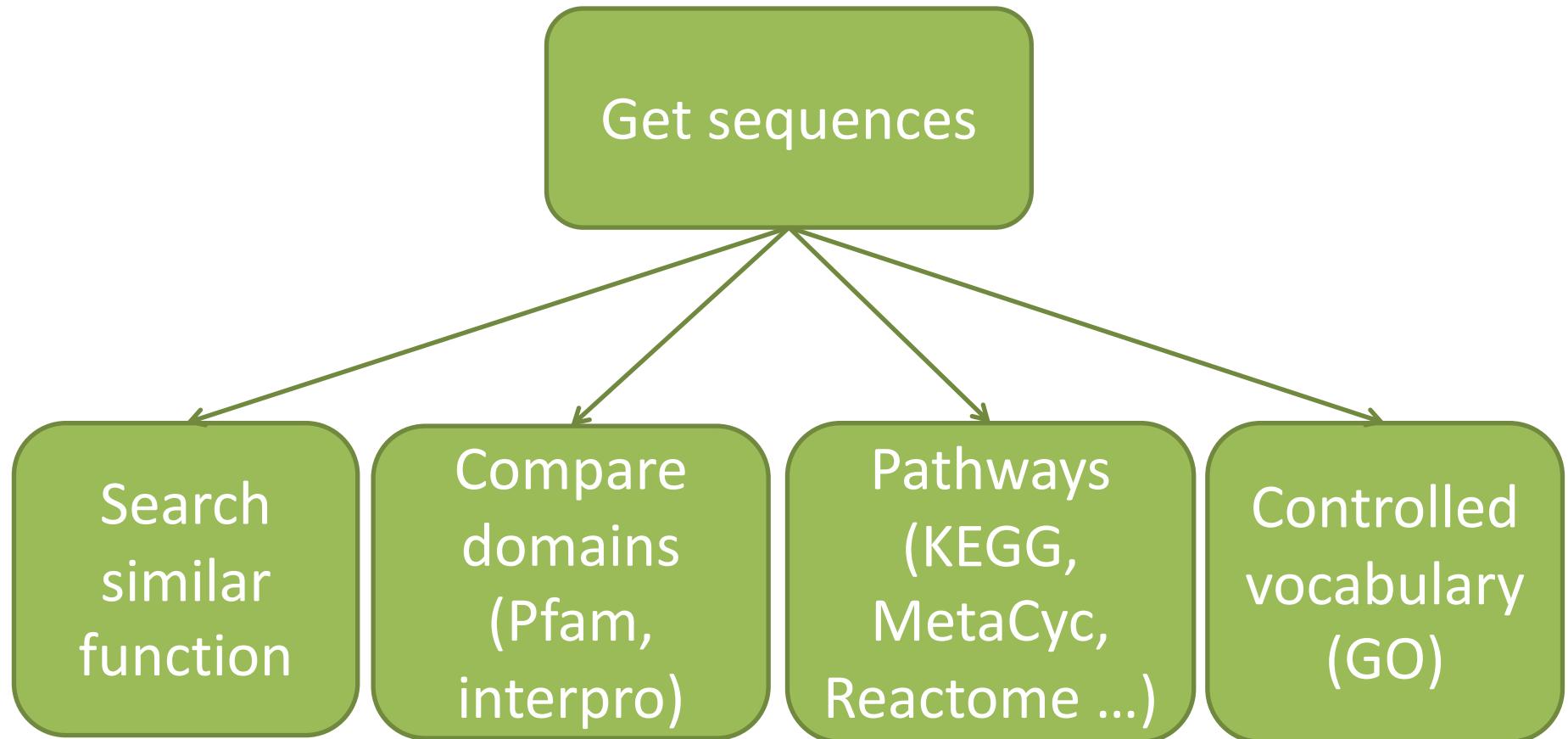
Blast-based annotation are tightly dependent to the quality of the structural annotation

- Gene Fusion
- Gene split
- Gene Partial (Well conserved domain)
- Over prediction
- Wrong ORF

Blast-based approach : results



Functional annotation – HOW?



Databases

Database	Information	Comment
KEGG	Pathway	Kyoto Encyclopedia of Genes and Genomes
MetaCyc	Pathway	Curated database of experimentally elucidated metabolic pathways from all domains of life (NIH)
Reactome	Pathway	Curated and peer reviewed pathway database
GO	Gene Ontology	Three structured, controlled vocabularies (ontologies) : biological processes, cellular components and molecular functions
Pfam	Protein families	Multiple sequence alignments and hidden Markov models
Interpro	Protein families, domains and functional sites	Run separate search applications, and create a signature to search against Interpro.
Have a look on the Interpro web page: All the database they search into are listed. It gives a nice overview of different types of databases available.		

Gene Ontology

Gene Ontology: the framework for the model of biology.

The GO defines concepts/classes used to describe gene function, and relationships between these concepts. It classifies functions along three aspects:

GO term prediction

Biological Process

- GO:0006631 fatty acid metabolic process
- GO:0006635 fatty acid beta-oxidation
- GO:0008152 metabolic process
- GO:0055114 oxidation-reduction process

Molecular Function

- GO:0003824 catalytic activity
- GO:0003857 3-hydroxyacyl-CoA dehydrogenase activity
- GO:0004300 enoyl-CoA hydratase activity
- GO:0016491 oxidoreductase activity
- GO:0016616 oxidoreductase activity, acting on the CH-OH group of donors, NAD or NADP as acceptor
- GO:0050662 coenzyme binding

Cellular Component

- GO:0005739 mitochondrion
- GO:0016507 mitochondrial fatty acid beta-oxidation multienzyme complex

More than 60 000 terms

pathways and larger processes
made up of the activities
of multiple gene products.

molecular activities
of gene products

where gene products are active

Gene Ontology

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



About Ontology Annotations Downloads Help



Current release 2019-05-09: 45,006 GO terms | 6,307,350 annotations
1,164,920 gene products | 4,455 species

THE GENE ONTOLOGY RESOURCE

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

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

Search GO term or Gene Product in AmiGO ...



Any Ontology Gene Product

GO Enrichment Analysis ?

Powered by PANTHER

Your gene IDs here...

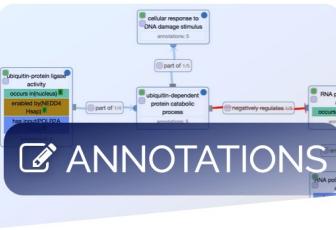
biological process

Homo sapiens

Examples

Launch ➔

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



ANNOTATIONS



TOOLS & GUIDES

Tools

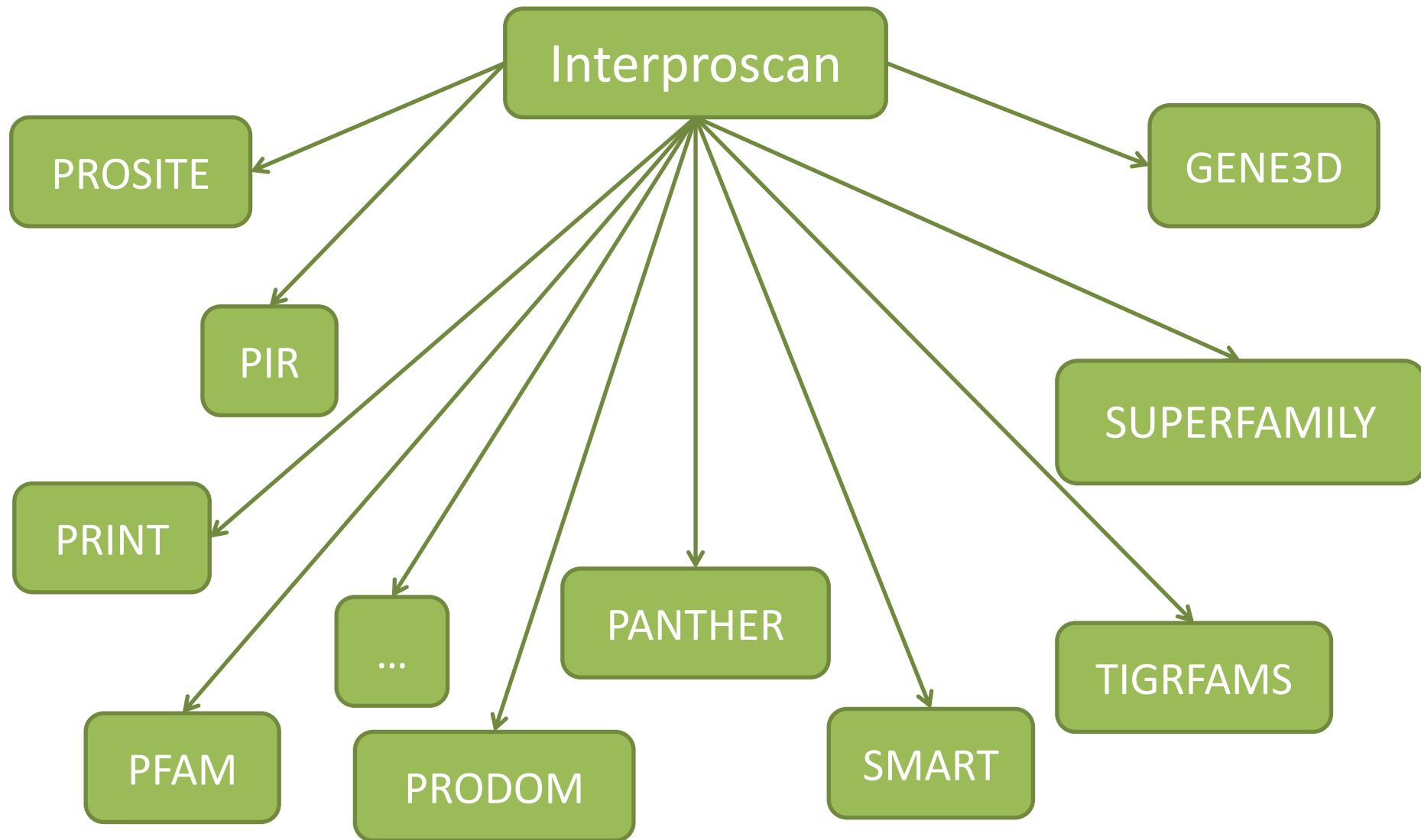
Tool	Approach	Comment
Trinotate	Best blast hit + protein domain identification (HMMER/PFAM) + protein signal peptide and transmembrane domain prediction (signalP/tmHMM), and leveraging various annotation databases (eggNOG/GO/Kegg databases).	Partially automated
Annocript	Best blast hit	Collects the best-hit and related annotations (proteins, domains, GO terms, Enzymes, pathways, short)
Annot8r	Best blast hits <u> </u>	A tool for Gene Ontology, KEGG biochemical pathways and Enzyme Commission EC number annotation of nucleotide and peptide sequences.
Sma3s	Best blast hit + Best reciprocal blast hit + clusterisation	3 annotation levels
afterParty	BLAST, InterProScan	web application
Interproscan	Run separate search applications HMMs, fingerprints, patterns => InterPro	Created to unite secondary databases
Blast2Go	Best* blast hits <u> </u>	Commercial !



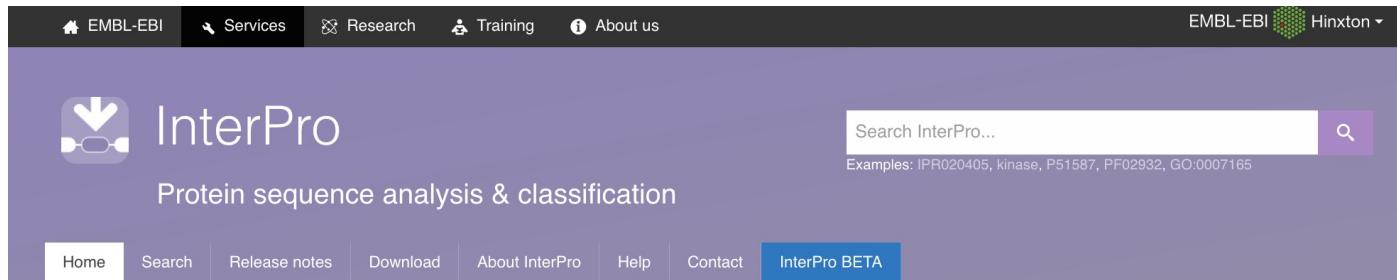
“InterPro is a resource that provides functional analysis of protein sequences by classifying them into families and predicting the presence of domains and important sites.

To classify proteins in this way, InterPro uses predictive models, known as signatures, provided by several different databases (referred to as member databases) that make up the InterPro consortium.”

<https://www.ebi.ac.uk/interpro/about.html>



- Annotate the sequences functionally using Interproscan : <http://www.ebi.ac.uk/interpro/>



InterPro: protein sequence analysis & classification

InterPro provides functional analysis of proteins by classifying them into families and predicting domains and important sites. We combine protein signatures from a number of member databases into a single searchable resource, capitalising on their individual strengths to produce a powerful integrated database and diagnostic tool. [Read more about InterPro](#)

Analyse your protein sequence

Submit | Clear | Example protein sequence

Documentation

[About InterPro](#): core concepts, update frequency, how to cite, team and consortium members.

[FAQs](#): what are entry types and why are they important, interpreting results, downloading InterPro?

[Web services documentation](#)

Protein focus

What's ape

The genus *Homo*, to which all human beings belong, is believed to have evolved from *Australopithecus* around 2–3 million years ago. Lucy, the *Australopithecus afarensis* ape, whose skeleton was pieced together from several hundred pieces of bone fossils, is the best known example of

Publications

[Nucleic Acids Research](#)
[InterPro in 2019: improving coverage, classification and access to protein sequence annotations](#)

Our latest paper describing new developments on the InterPro website (*Nucleic Acids Research*, Jan 2019).
[HTML](#) | [PDF \(5.7Mb\)](#) | [All publications](#)

v74 InterPro 74.0
9th May 2019

Features include:

- The addition of 156 InterPro entries.
- Integration of 174 new methods from the CATH-Gene3D (100), CDD (28) and PANTHER (46) databases.
- Removal of the ProDom (2006.1) database.

[Download](#) | [Read more](#)

Tweets by @InterProDB

 InterPro
@InterProDB

Replying to @InterProDB

InterProScan 5 (version 5.34-73.0) is now available! For more details please visit: [github.com/ebi-pf-team/interproscan...](https://github.com/ebi-pf-team/interproscan)

 ebi-pf-team/interproscan
Contribute to ebi-pf-team/...
github.com

Mar 28, 2019

 InterPro



Contents and coverage of InterPro 74.0

InterPro protein matches are now calculated for all UniProtKB and UniParc proteins. The following statistics are for all UniProtKB proteins.

InterPro release 74.0 contains [36713](#) entries (last entry: [IPR042311](#)), representing:

 Homologous superfamily (3078)

 Family (21769)

 Domain (10637)

 Repeat (316)

 Sites

... Active site (132)

... Binding site (76)

... Conserved site (688)

... PTM (17)

InterPro cites 58657 publications in PubM

 Structural domains

Member database information

Signature database	Version	Signatures*	Integrated signatures**
CATH-Gene3D	4.2.0	6119	2369
CDD	3.16	12805	3284
HAMAP	2019_01	2274	2245
PANTHER	14.1	123151	9043
Pfam	32.0	17929	17421
PIRSF	3.02	3285	3217
PRINTS	42.0	2106	1953
PROSITE patterns	2019_01	1310	1287
PROSITE profiles	2019_01	1232	1173
SFLD	4	303	147
SMART	7.1	1312	1264
SUPERFAMILY	1.75	2019	1601
TIGRFAMs	15.0	4488	4435

* Some signatures may not have matches to UniProtKB proteins.

** Not all signatures of a member database may be integrated at the time of an InterPro release

Other sequence features

Coils Phobius SignalP TMHMM

Interproscan



Sequence database		Version		Count of proteins matching
			any signature	integrated signatures
UniProtKB	2019_04	156637804	130888307 (83.6%)	126806860 (81.0%)
UniProtKB/TrEMBL	2019_04	156077686	130343729 (83.5%)	126265196 (80.9%)
UniProtKB/Swiss-Prot	2019_04	560118	544578 (97.2%)	541664 (96.7%)

InterPro2GO

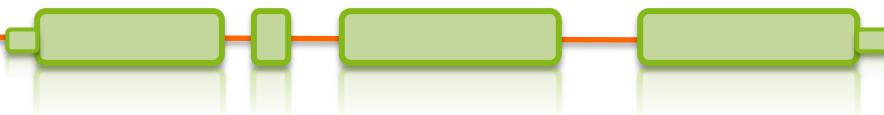
Total number of GO terms mapped to InterPro entries - 34141

Not integrated signatures = signature not yet curated or do not reach InterPro's standards for integration

pathway information available as well:

- KEGG
- MetaCyc
- Reactome
- UniPathway

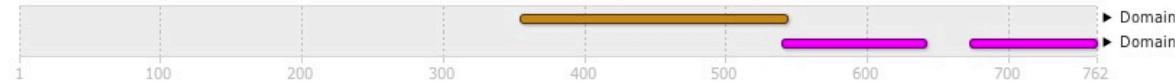
Interproscan results



Protein family membership

- Crotonase superfamily (IPR001753)
- Fatty acid oxidation complex, alpha subunit, mitochondrial (IPR012803)

Domains and repeats



Detailed signature matches



Output: TSV, XML, SVG, etc

gene-2.44-mRNA-1 a9deba5837e2614a850c7849c85c8e9c 447 Pfam PF02458 Transferase family 98 425
1.4E-15 T 31-10-2015 IPR003480 Transferase GO:0016747

gene-0.13-mRNA-1 61882f1a46b15c8497ed9584a0eb1a35 459 Pfam PF01490 Transmembrane amino acid
transporter protein 49 439 2.0E-39 T 31-10-2015 IPR013057 Amino acid transporter, transmembrane

gene-1.4-mRNA-1 b867bbb377084bba6ea84dcda9f27f4e 511 SUPERFAMILY SSF103473 42 481
4.19E-50 T 31-10-2015 IPR016196 Major facilitator superfamily domain, general substrate transporter

gene-1.4-mRNA-1 b867bbb377084bba6ea84dcda9f27f4e 511 Pfam PF07690 Major Facilitator Superfamily 67
447 3.5E-30 T 31-10-2015 IPR011701 Major facilitator superfamily GO:0016021|GO:0055085

Scripts exist to merge the interproscan-results to the structural annotation gff file



Another way : use the (mostly) commercial alternative



- Combines a blast-based search with a search for functional domains
- Blast at NCBI -> picks out GO terms based on blast hits and uniprot -> statistical significance test -> done!
- Blast2Go relies entirely on sequence similarity ... but InterProScan searches can also be launched within blast2go
- Command line tool or Plugin for Geneious or CLC bio Workbench (commercial tools for downstream analyses)

=> Contain nice downstream analysis/visualization components

Blast2GO V.2.4.4

File Blast Mapping Annotation Analysis Statistics Select Tools View Info

GO:0007067,GO:0016021 transport;binding;apoptos SPO_2518,DDX18_HUMAN

nr	sequence name	seq description	length	#...	min. eValue	sim mean	#G...	GO IDs	Enzyme	InterPro
	c6 transcription		977	20	1.0E-171	59.85%	7	Etranscription factor activity; Ezinc ion binding; Pregulation of transcription, DNA-dependent; Ctranscription factor complex; Etransporter activity; Cmembrane; Ptransmembrane transport		IPRO05829; IPRO07219
3884	gene_3884 GeneMark...							Cviral capsid		no IPS match
3885	gene_3885 GeneMark...	hypothetical protein NFIA_039100 [Neosartorya fischeri NRRL 181]	312	20	1.0E-39	63.15%	1			-
3886	gene_3886 GeneMark...	sin3 complex subunit	870	20	0.0	73.2%	0			-
3887	gene_3887 GeneMark...	mitochondrial intermembrane space translocase subunit	87	20	1.0E-40	88.55%	5	Emetal ion binding; Pprotein import into mitochondrial inner membrane; Cmitochondrial inner membrane; Cmitochondrial intermembrane space protein transporter complex, Ptransmembrane transport		IPRO04217; PTHR11038 (PANTHER), PTHR11038_SF8 (PANTHER)
		lysyl-tRNA synthetase						Ccytoplasm; Pauxin biosynthetic process; Enucleic acid binding; Elysine-tRNA ligase activity; Plysyl-tRNA aminoacylation; EATP binding; Plysine biosynthetic process		IPRO04364; IPRO04365; IPRO06195; IPRO12340; IPRO16027; IPRO18149; IPRO18150; G3DSA:3.30.930.10 (GENE3D), SSF5568 (SUPERFAMILY)
3888	gene_3888 GeneMark...		592	20	0.0	73.55%	7		EC:6.1.1.6	
3889	gene_3889 GeneMark...	transcription factor conserved	1569	20	0.0	70.9%	0			-
3890	gene_3890 GeneMark...	hypothetical protein [Aspergillus clavatus NRRL 1]	240	20	1.0E-51	56.25%	0			-
		udp-glc gal endoplasmic reticulum nucleotide						Cintegral to membrane; Cendoplasmic reticulum membrane; Ptransmembrane transport; Pcarbohydrate transport		IPRO13657; PTHR10778 (PANTHER)

GO Graphs Application Messages Blast/IPS Results Statistics Kegg Maps

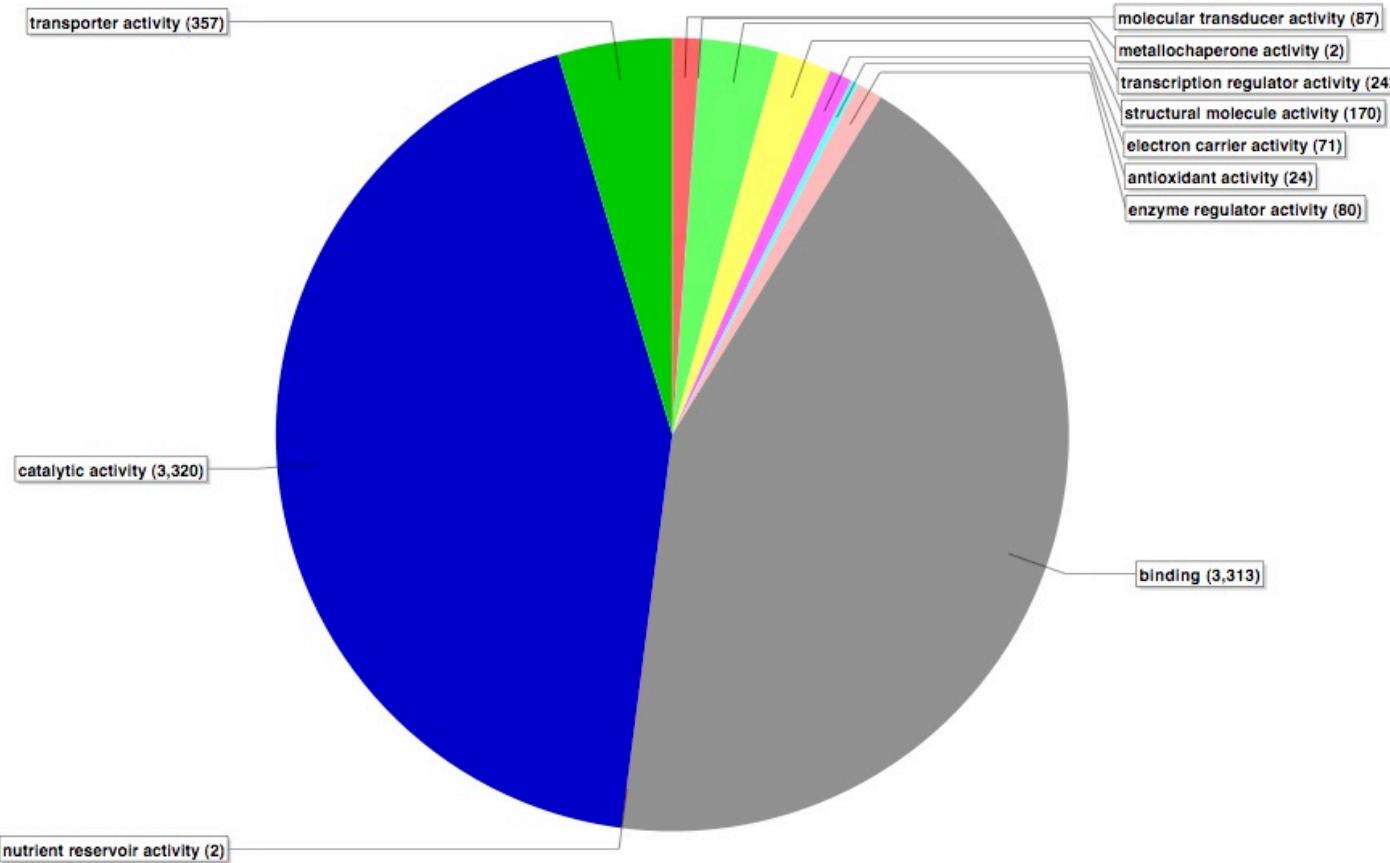
```

17:59 InterProScan for gene_8871|GeneMark.hmm|286_aa done.
17:59 -----
17:59 InterProScan Result:
17:59 InterProID: IPRO01715
17:59 InterProName: Calponin-like actin-binding
17:59 InterProType: Domain
17:59 DB-Name: GENE3D - G3DSA:1.10.418.10
17:59 InterProID: IPRO16146
17:59 InterProName: Calponin-homology
17:59 InterProType: Domain
17:59 DB-Name: SUPERFAMILY - SSF47576
17:59 InterProID: noIPR
17:59 InterProName: unintegrated
17:59 InterProType: unintegrated
17:59 DB-Name: PANTHER - PTHR19961
17:59 DB-Name: PANTHER - PTHR19961:SF9

```

Annotation already running



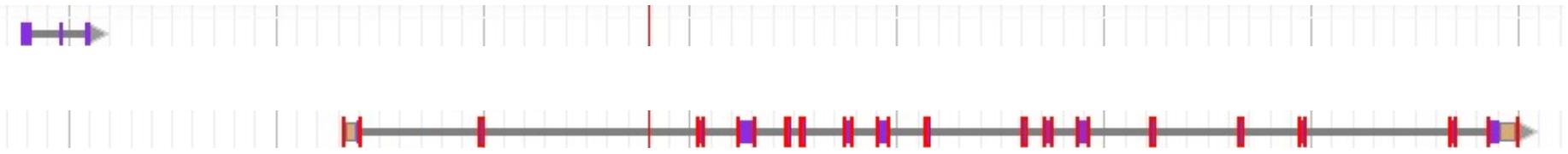


Quick view of synteny-based method



Liftovers are very useful for orthology determination

- Align two genomes (Satsuma) (<http://satsuma.sourceforge.net/>)
- Transfer annotations between aligned regions (Kraken)(<https://github.com/nedaz/kraken>)
- Transfer functional annotations between lifted genes that overlap annotated genes



One word about network

Categorizations of gene function (e.g GO) in a hierarchy of categories is helpful

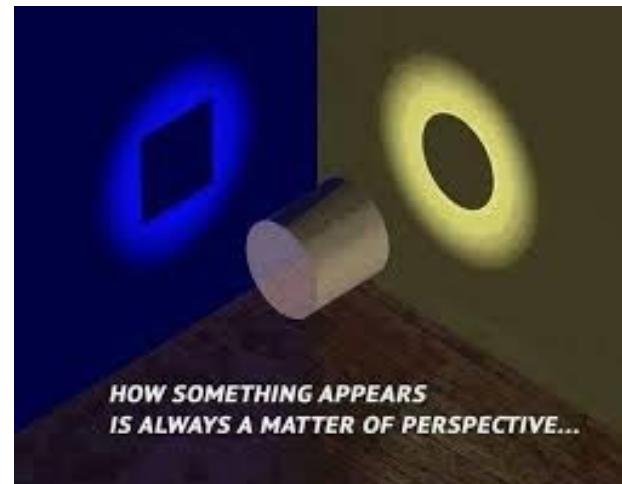
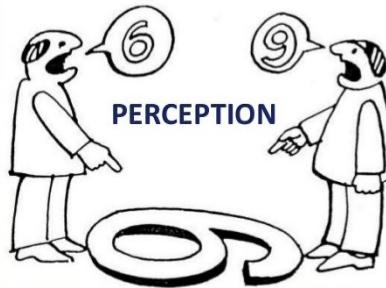
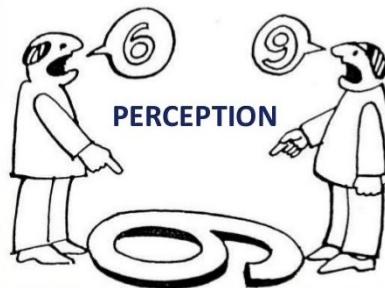
BUT

gene has no function alone

⇒ Pathways / regulatory networks explain how genes interact so what they are doing!

E.g. databases for pathway :

- KEGG
- MetaCyc
- Reactome
- UniPathway



KEGG-mapping

File Blast Mapping Annotation Analysis Statistics Select Tools View Help

GO:0007067, GO:0016021 transport;binding;apoptos SPO_2518, DDX18_HUMAN

Enzyme IPR003781; IPR005810

GO Graphs Application Messages Blast/IPS Results Statistics Kegg Maps

GLYCEROLIPID METABOLISM

Pathways

- Pentose phosphate pathway
- Fructose and mannose metabolism
- Butanoate metabolism
- Carbon fixation in photosynthetic organisms
- Lysine degradation
- Tyrosine metabolism
- Methane metabolism
- Glyoxylate and dicarboxylate metabolism
- Glycerolipid metabolism**
- Glutathione metabolism
- Selenoamino acid metabolism
- Phenylalanine metabolism
- Benzoate degradation via CoA ligation
- Valine, leucine and isoleucine biosynthesis
- Reductive carboxylate cycle (CO₂ fixation)
- Galactose metabolism
- Phenylalanine, tyrosine and tryptophan biosynthesis
- N-Glycan biosynthesis
- Photosynthesis
- Drug metabolism – other enzymes
- Sulfur metabolism
- Fatty acid biosynthesis
- Inositol phosphate metabolism
- beta-Alanine metabolism
- Drug metabolism – cytochrome P450
- Pantothenate and CoA biosynthesis
- Biosynthesis of unsaturated fatty acids
- Cyanoamino acid metabolism
- Terpenoid backbone biosynthesis
- Histidine metabolism
- T cell receptor signaling pathway
- Tropamine, piperidine and pyridine alkaloid biosynthesis
- One carbon pool by folate
- Pentose and glucuronate interconversions
- Phosphatidylinositol signaling system

Color	Enzyme	Sequences
red	ec:1.1.1.2 - alcohol dehydrogenase (NADP+)	gene_674 GeneMark.hmm 333_aa, gene_5801 GeneMark.hmm 312_aa
yellow	ec:2.3.1.158 - phospholipid:diacylglycerol acyltransferase	gene_2604 GeneMark.hmm 188_aa, gene_6532 GeneMark.hmm 505_aa
orange	ec:2.3.1.51 - 1-acylglycerol-3-phosphate O-acyltransferase	gene_176 GeneMark.hmm 429_aa, gene_6693 GeneMark.hmm 292_aa
green	ec:2.3.1.20 - diacylglycerol O-acyltransferase	gene_176 GeneMark.hmm 429_aa, gene_7213 GeneMark.hmm 521_aa, gene_8170 GeneMark.hmm 470_aa
blue	ec:2.3.1.15 - glycerol-3-phosphate O-acyltransferase	gene_886 GeneMark.hmm 748_aa, gene_2640 GeneMark.hmm 823_aa
pink	ec:1.1.1.72 - glycerol dehydrogenase (NADP+)	gene_3376 GeneMark.hmm 325_aa, gene_4577 GeneMark.hmm 326_aa
violet	ec:1.2.1.3 - aldehyde dehydrogenase (NAD+)	gene_2201 GeneMark.hmm 497_aa, gene_5247 GeneMark.hmm 502_aa, gene_5611 GeneMark.hmm 471_aa
light-red	ec:2.7.1.107 - diacylglycerol kinase	gene_5292 GeneMark.hmm 409_aa

Annotation already running

- Functional annotation found
 /!\ Transmission of error from databases !
 Experimental check is good !
 - Hypothetical protein / Uncharacterized protein
 => depends largely on conventional experiments.
- Knowing the function is not enough: Chimp and human => 98% similarity
 => Knowledge of other parameters useful (pathway, positional and temporal regulation of genes)