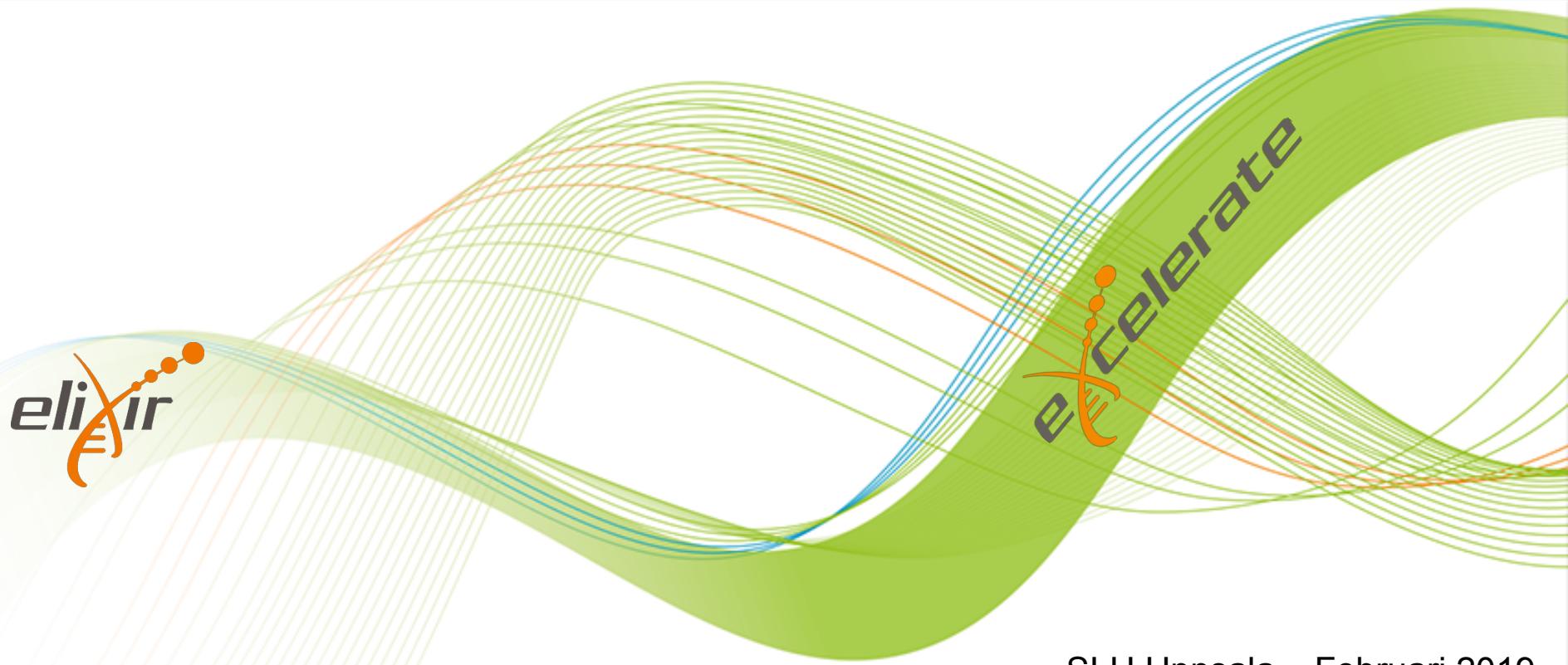
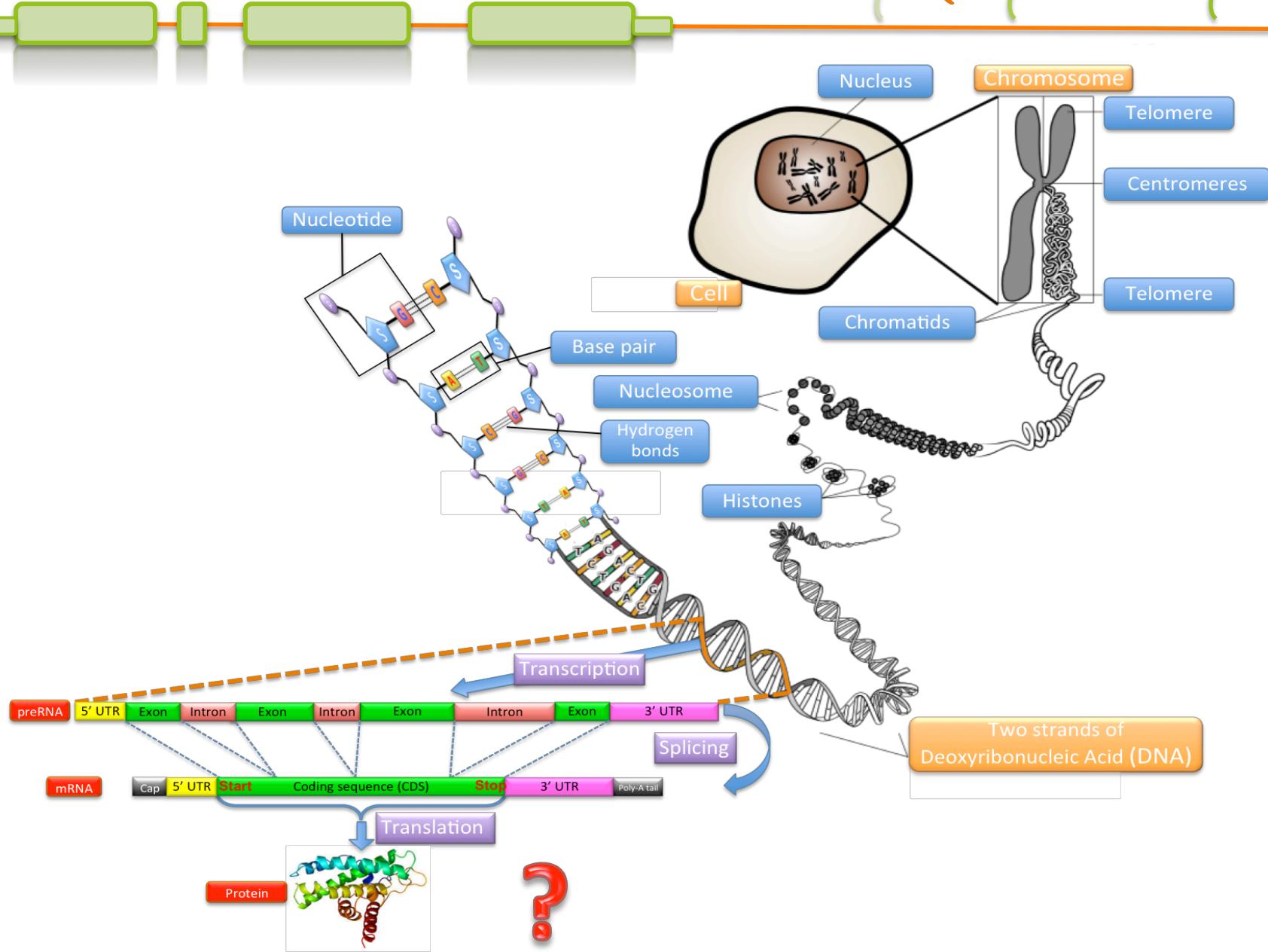


Jacques Dainat PhD

Functional annotation



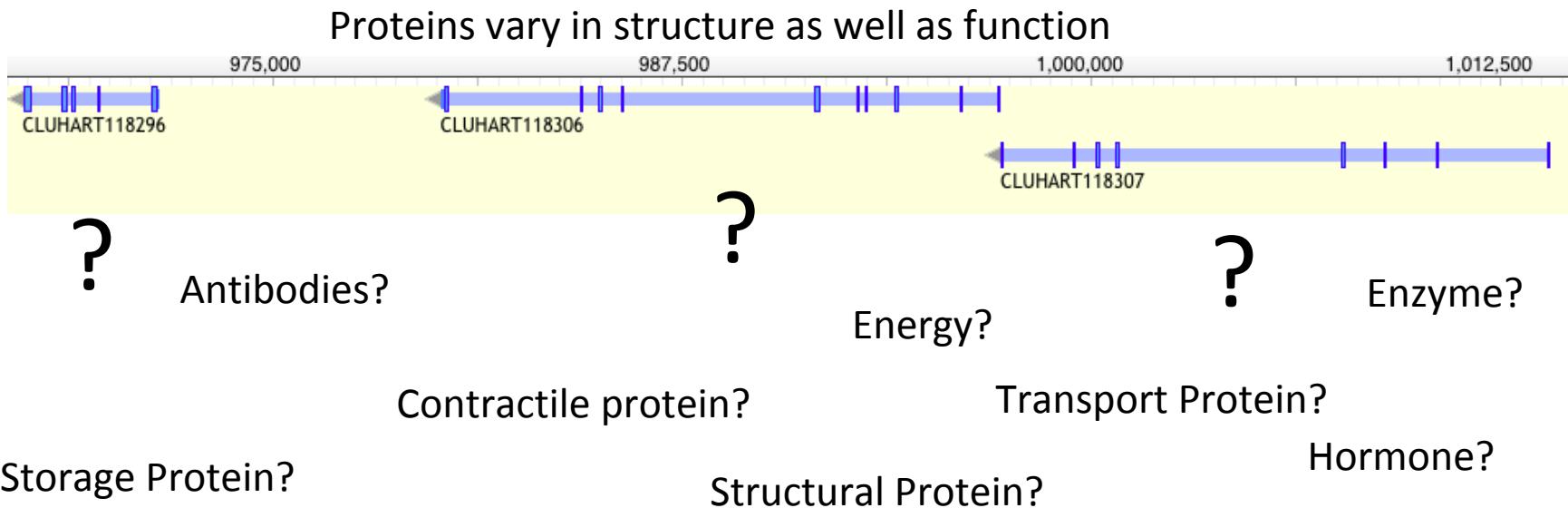
Overview



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?

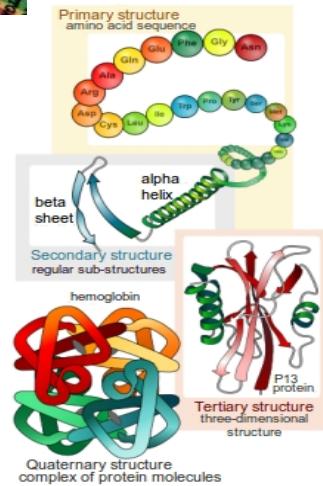
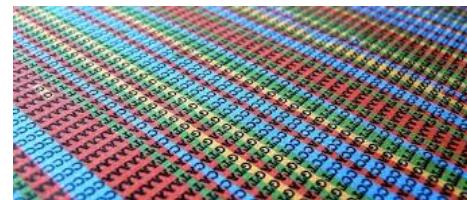


- Experimentally
 - => Mutants, knockout, etc.
- Accurate

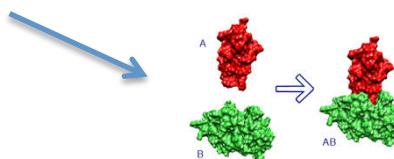


Mice homozygous for the diabetes 3J spontaneous mutation

- Computationally
 - Sequence-based
 - Structure based
 - Protein-protein interaction data



limited accuracy



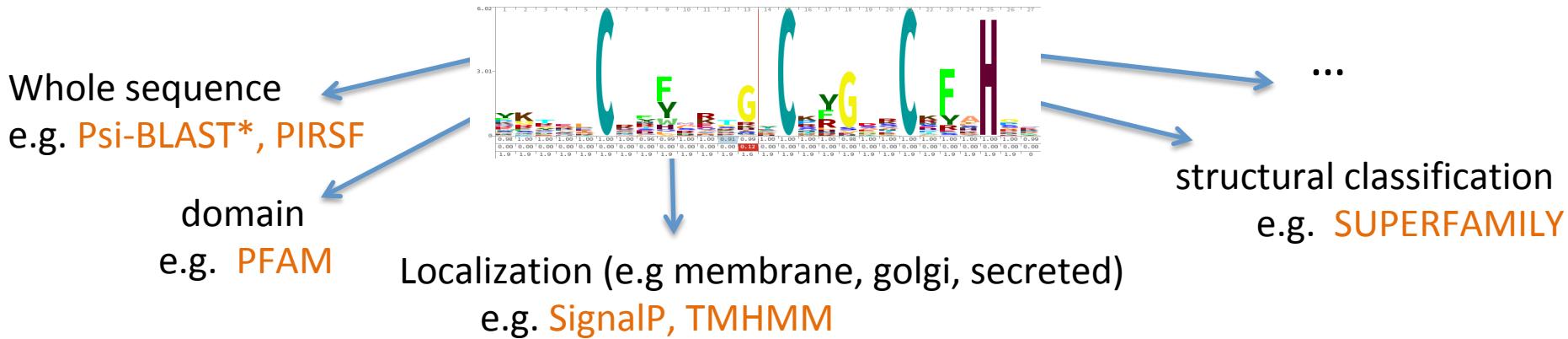
Methods - Sequence-based

- Based on similarity
=>Best blast hit

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

- Based on Motif
⇒Proscan, MEME, QuasiMotifFinder
- Based on Profile (HMM or other statistical signature)

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



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

- **Similarity to known structures.**

- Global structure-comparison
 - CATH and SCOP, the two most comprehensive structure-based family resources
- 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 for your protein?

=> Importance to gathering 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?

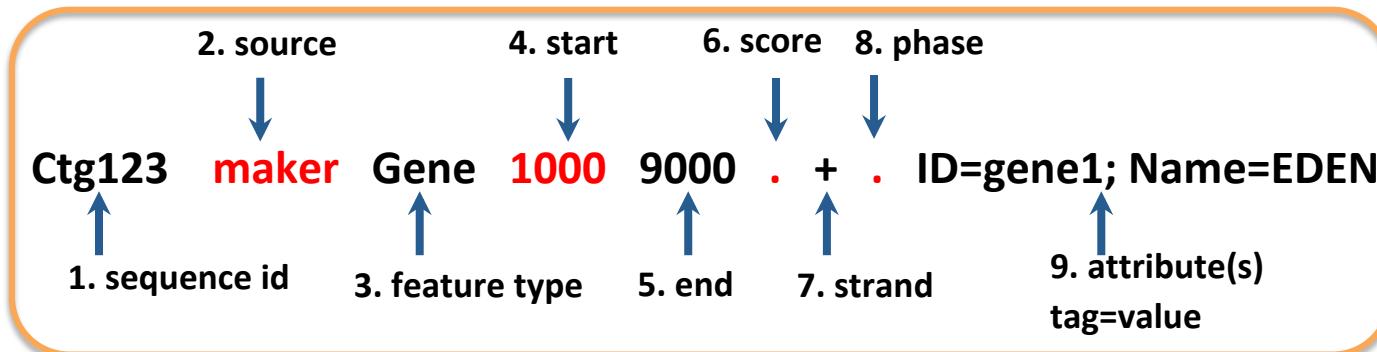


Get sequences

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

Blast-based approach

Annotate the sequences functionally using Blast

Choose
database

Uniprot	Swissprot
exhaustive	reliable

Blast-based approach

Annotate the sequences functionally using Blast

Choose
database

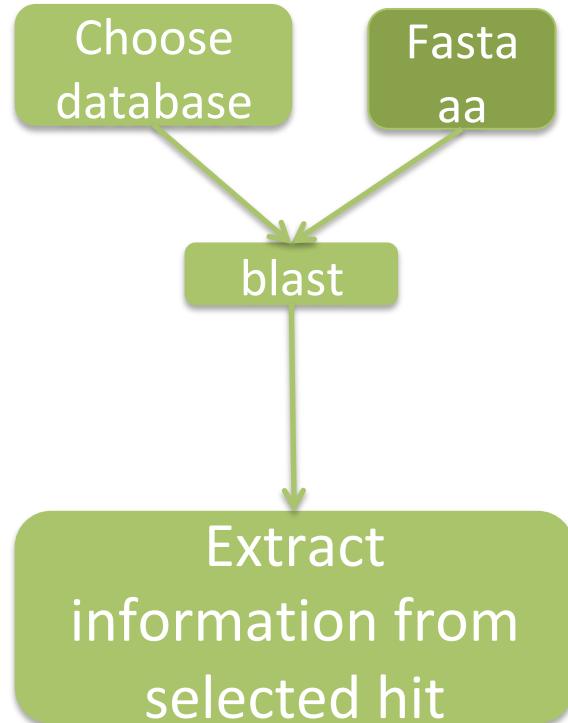
Fasta
aa

blast



Minimum Threshold

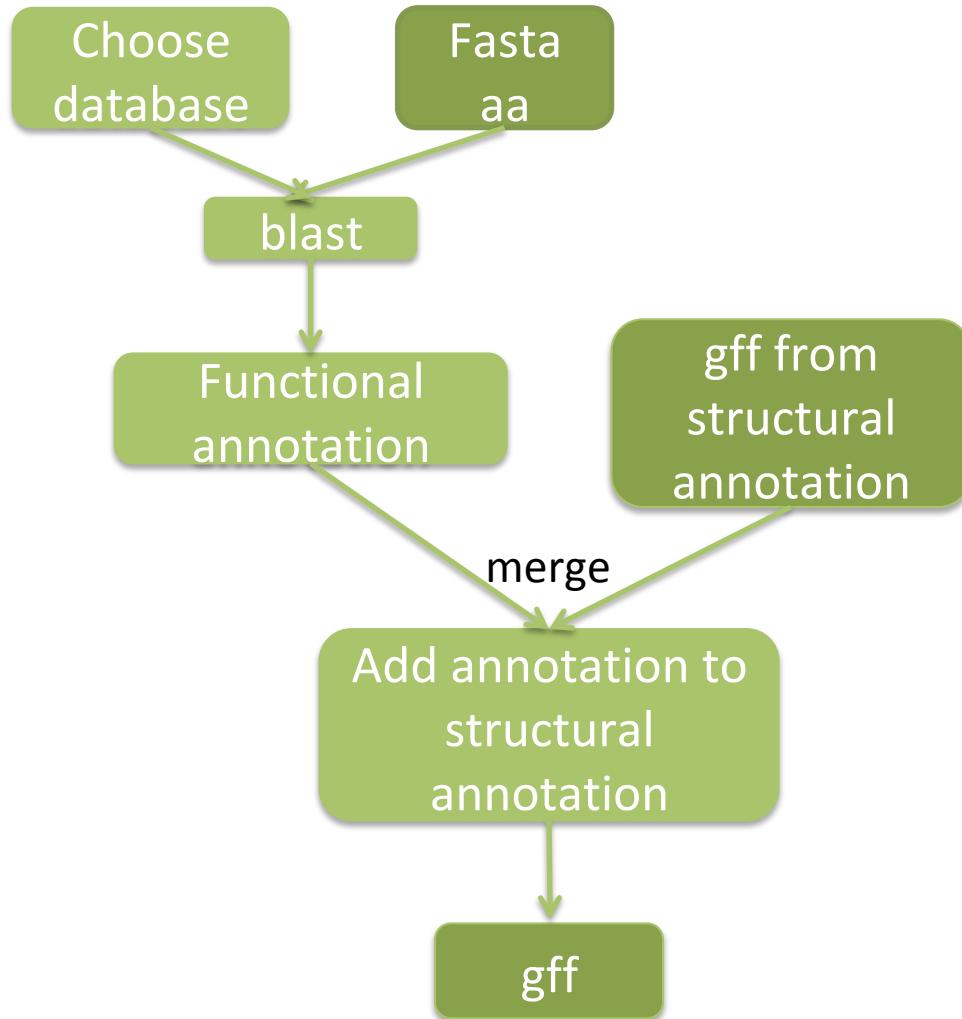
Annotate the sequences functionally using Blast



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

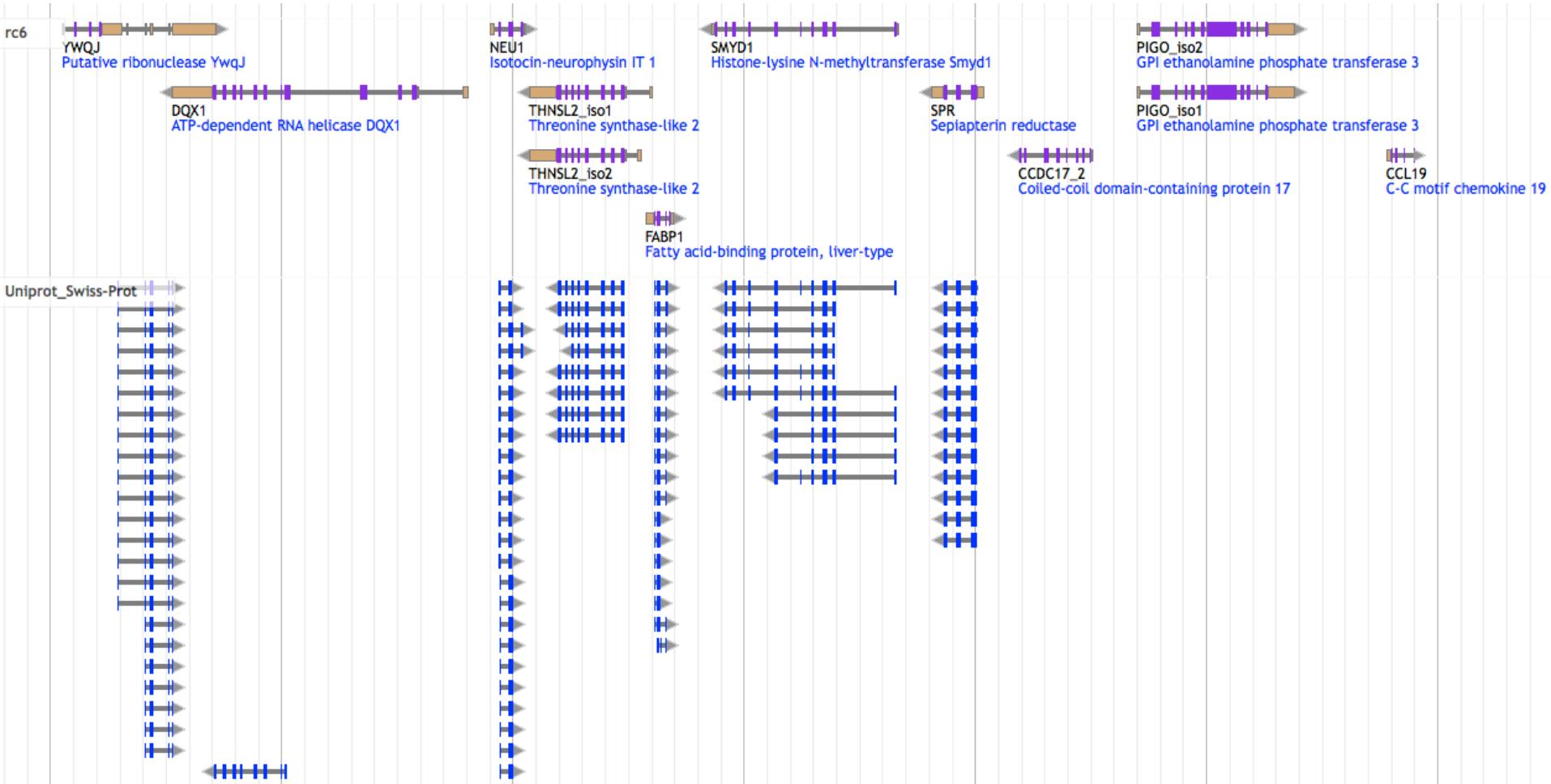
- Gene Fusion
- Gene split
- Gene Partial (Well conserved domain)
- Over prediction
- Wrong ORF

Blast-based approach : result

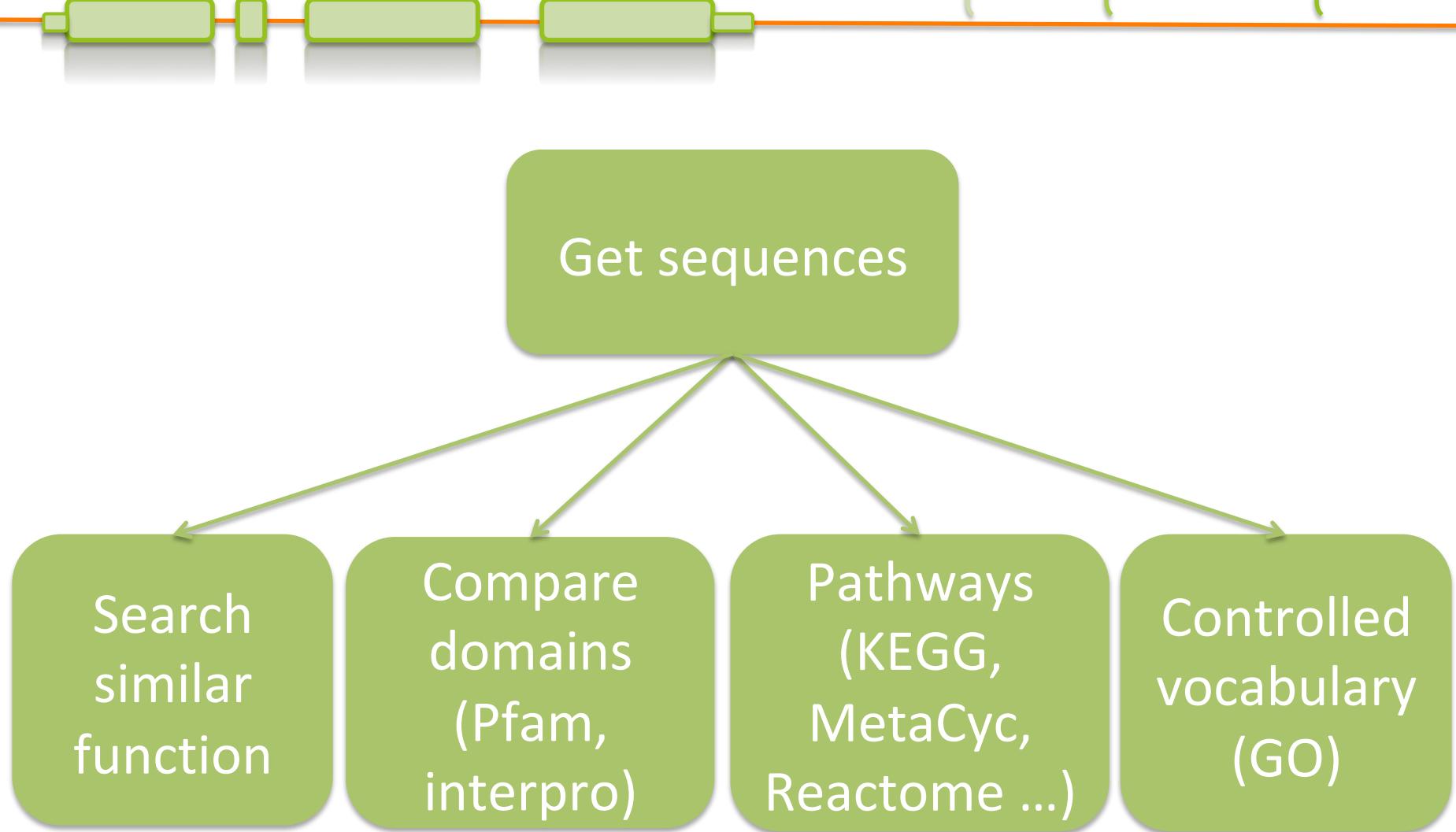


SciLifeLab

NBIS
NATIONAL BIOINFORMATICS
INFRASTRUCTURE SWEDEN



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
UniPathway	Pathway	Manually curated resource of enzyme-catalyzed and spontaneous chemical reactions.
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

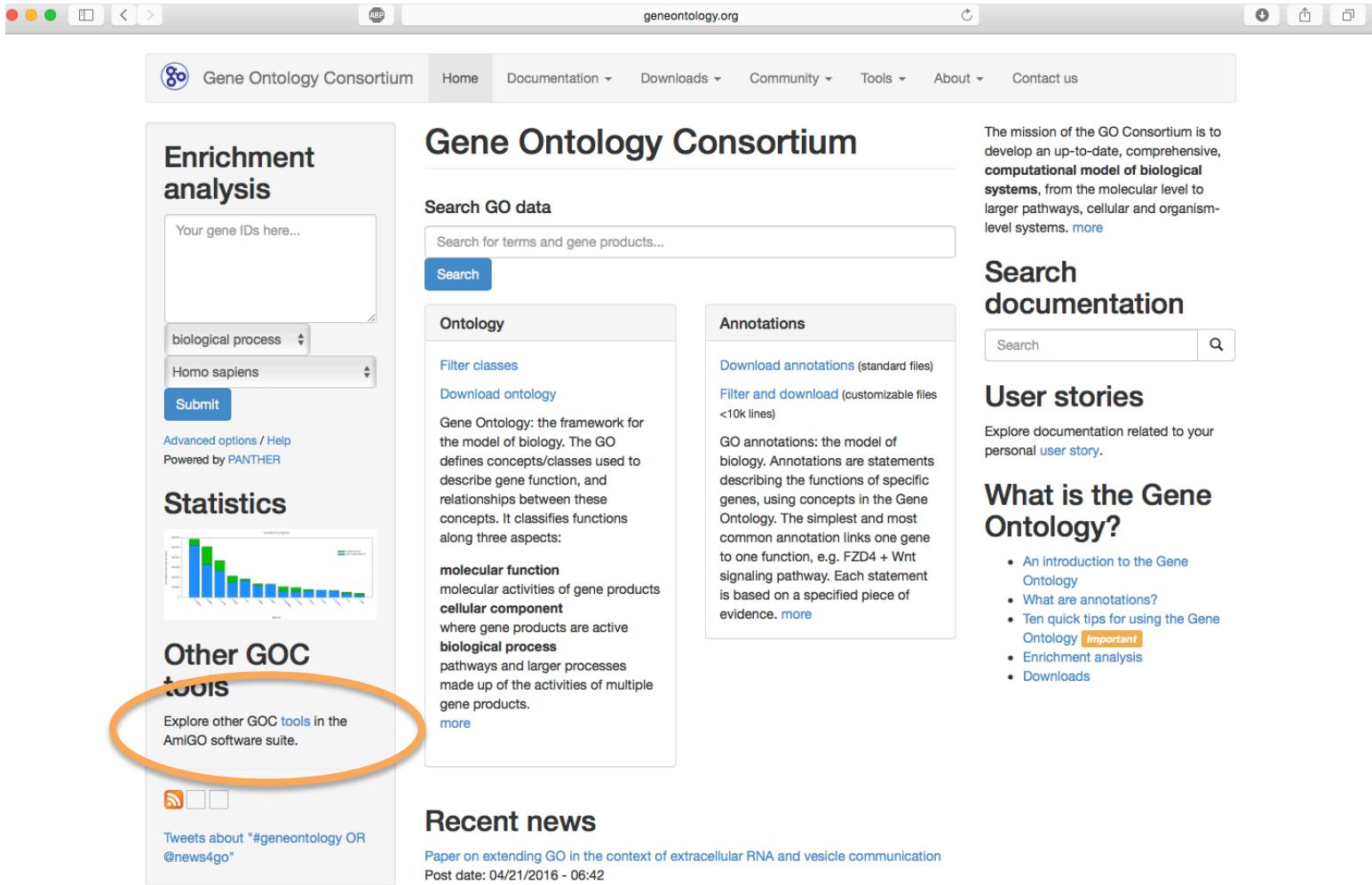
pathways and larger processes made up of the activities of multiple gene products.

molecular activities of gene products

where gene products are active

Gene Ontology

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



The screenshot shows the homepage of the Gene Ontology Consortium. At the top, there's a navigation bar with links for Home, Documentation, Downloads, Community, Tools, About, and Contact us. Below the navigation is a main content area.

Enrichment analysis section (left sidebar):

- Form for "Your gene IDs here..."
- Dropdown menu: biological process
- Dropdown menu: Homo sapiens
- Submit button
- Advanced options / Help
- Powered by PANTHER

Statistics section:

- Bar chart showing some data across categories.

Other GOC tools section (circled in orange):

- Text: Explore other GOC tools in the AmiGO software suite.
- Social media icons: RSS, GitHub, Bitbucket.

Tweets about #geneontology OR @news4go section:

Gene Ontology Consortium section (main header):

Gene Ontology Consortium

Search GO data

Search for terms and gene products...

Ontology

- Filter classes
- Download 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:

- molecular function
- molecular activities of gene products
- cellular component
- where gene products are active
- biological process
- pathways and larger processes made up of the activities of multiple gene products.

[more](#)

Annotations

- Download annotations (standard files)
- Filter and download (customizable files <10k lines)

GO annotations: the model of biology. Annotations are statements describing the functions of specific genes, using concepts in the Gene Ontology. The simplest and most common annotation links one gene to one function, e.g. FZD4 + Wnt signaling pathway. Each statement is based on a specified piece of evidence. [more](#)

Search documentation

Search

User stories

Explore documentation related to your personal [user story](#).

What is the Gene Ontology?

- An introduction to the Gene Ontology
- What are annotations?
- Ten quick tips for using the Gene Ontology Important
- Enrichment analysis
- Downloads

Recent news

Paper on extending GO in the context of extracellular RNA and vesicle communication
Post date: 04/21/2016 - 06:42

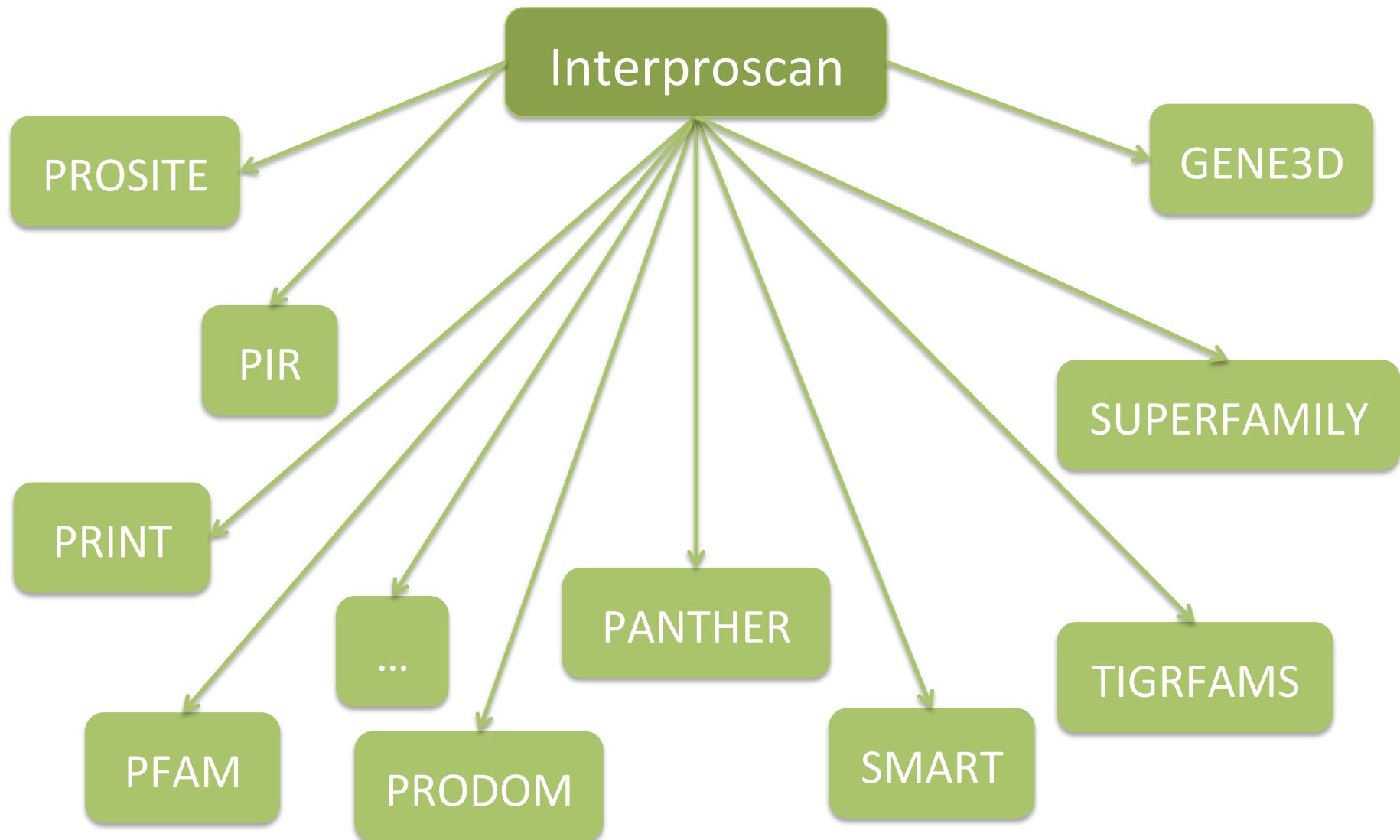
Tools

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).	Not automated
Annocrypt	Best blast hit	Collects the best-hit and related annotations (proteins, domains, GO terms, Enzymes, pathways, short)
Annot8r	Best blast hits	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	Retrieve only GO Commercial !

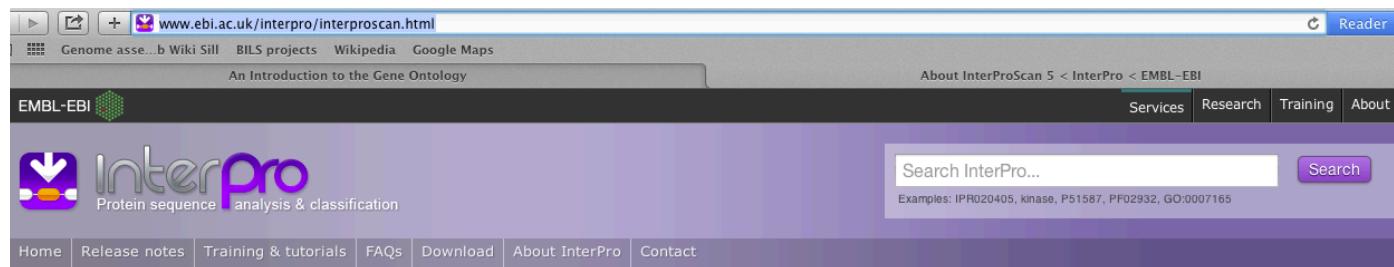
“InterPro is a resource that provides functional analysis of protein sequences by classifying them into families and predicting the presence of domains and important sites.

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



The screenshot shows the EBI InterPro homepage. At the top, there's a navigation bar with links like 'Home', 'Release notes', 'Training & tutorials', 'FAQs', 'Download', 'About InterPro', and 'Contact'. Below the navigation is a search bar with the placeholder 'Search InterPro...' and a 'Search' button. To the right of the search bar, it says 'Examples: IPR020405, kinase, P51587, PF02932, GO:0007165'. The main content area features the 'InterPro' logo and the text 'Protein sequence analysis & classification'. There are also links for 'Services', 'Research', 'Training', and 'About us'.

About InterProScan

What is InterProScan?

InterProScan is the software package that allows sequences (protein and nucleic) to be scanned against InterPro's signatures. Signatures are predictive models, provided by several different databases (referred to as member databases), that make up the InterPro consortium.

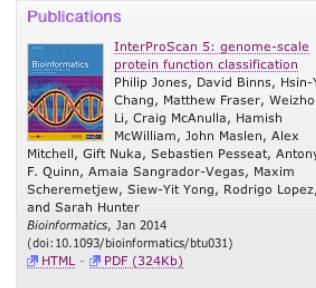
The software is available:

- As a web-based tool, using the sequence search box on the [InterPro homepage](#), for the analysis of single protein sequences (also available in the [EBI tool section](#))
- Programmatically via Web services that allow up to 25 sequences to be analysed per request (both [SOAP](#) and [REST](#)-based services are available)
- As a downloadable package for local installation from the EBI's FTP server, for instructions see the [detailed documentation pages](#).

InterProScan is run regularly against UniProtKB and the results are made available via the InterPro website.

More information

For more information, and for instructions on how to obtain, install and run InterProScan, please see the [detailed documentation pages](#).

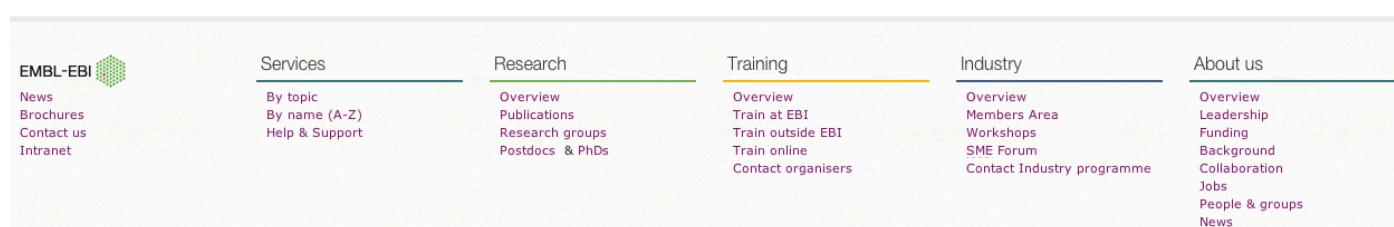


Publications

InterProScan 5: genome-scale protein function classification
Philip Jones, David Binns, Hsin-Yu Chang, Matthew Fraser, Weizhong Li, Craig McAnulla, Hamish McWilliam, John Maslen, Alex Mitchell, Giff Nuka, Sébastien Pesquet, Antony F. Quinn, Amaia Sangrador-Vegas, Maxim Scheremetjew, Siew-Yit Yong, Rodrigo Lopez, and Sarah Hunter
Bioinformatics, Jan 2014
(doi:10.1093/bioinformatics/btu031)
[HTML](#) - [PDF \(324Kb\)](#)

Jones,P.et.al.InterProScan5:genome-scale protein function classification.
Bioinformatics 30, 1236–1240 (2014).

Quevillon E., Silventoinen V., Pillai S., Harte N., Mulder N., Apweiler R., et al. . (2005). InterProScan: protein domains identifier. *Nucleic Acids Res.* 33, W116–W120. 10.1093/nar/gki442



EMBL-EBI

[News](#)
[Brochures](#)
[Contact us](#)
[Intranet](#)

Services

[By topic](#)
[By name \(A-Z\)](#)
[Help & Support](#)

Research

[Overview](#)
[Publications](#)
[Research groups](#)
[Postdocs & PhDs](#)

Training

[Overview](#)
[Train at EBI](#)
[Train outside EBI](#)
[Train online](#)
[Contact organisers](#)

Industry

[Overview](#)
[Members Area](#)
[Workshops](#)
[SME Forum](#)
[Contact Industry programme](#)

About us

[Overview](#)
[Leadership](#)
[Funding](#)
[Background](#)
[Collaboration](#)
[Jobs](#)
[People & groups](#)
[News](#)

Contents and coverage of InterPro 62.0

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

InterPro release 62.0 contains [29930](#) entries (last entry: [IPR034768](#)), representing:

 Family (19869)

 Domain (8868)

 Repeat (282)

 Sites

 ↳ Active site (132)

 ↳ Binding site (76)

 ↳ Conserved site (686)

 ↳ PTM (17)

InterPro cites 51421 publications in PubMed.

 Structural domains



Member database information

Signature database	Version	Signatures*	Integrated signatures**
CATH-Gene3D	4.1.0	2737	1198
CDD	3.14	11273	1526
HAMAP	201701.18	2160	2160
PANTHER	11.1	91538	5923
Pfam	30.0	16306	15710
PIRSF	3.01	3285	3222
PRINTS	42.0	2106	1986
ProDom	2006.1	1894	1131
PROSITE patterns	20.132	1309	1289
PROSITE profiles	20.132	1174	1142
SFLD	2	480	146
SMART	7.1	1312	1265
SUPERFAMILY	1.75	2019	1461
TIGRFAMs	15.0	4488	4450

* 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	Count of proteins matching	
			any signature	integrated signatures
UniProtKB	2017_03	80758400	71118703 (88.1%)	64919649 (80.4%)
UniProtKB/TrEMBL	2017_03	80204459	70576370 (88.0%)	64384952 (80.3%)
UniProtKB/Swiss-Prot	2017_03	553941	542333 (97.9%)	534697 (96.5%)

InterPro2GO

Total number of GO terms mapped to InterPro entries - 32178

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

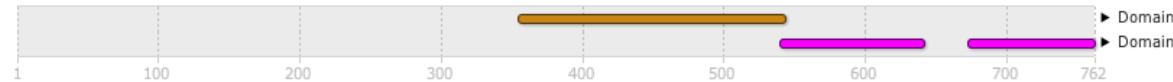
Interproscan results



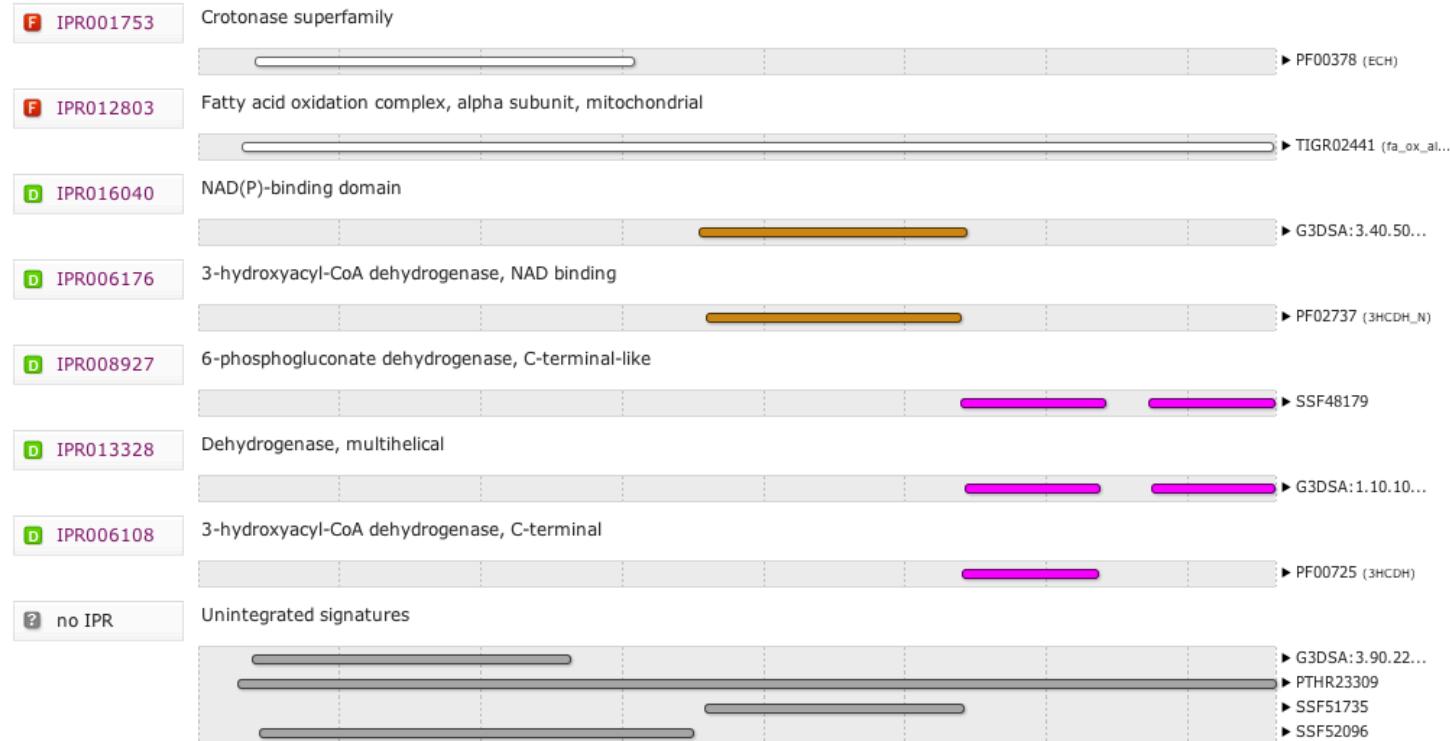
Protein family membership

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

Domains and repeats



Detailed signature matches



Ouput: 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

/Users/hobbe/Documents/Artemis_files_current/blast2go_20101001_0816.dat - 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
	3884	gene_3884 GeneMark...	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
	3885	gene_3885 GeneMark...	hypothetical protein NFIA_039100 [Neosartorya fischeri NRRL 181]	312	20	1.0E-39	63.15%	1	Cviral capsid	-	no IPS match
	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)
	3888	gene_3888 GeneMark...	lysyl-tRNA synthetase	592	20	0.0	73.55%	7	Ccytoplasm; Pauxin biosynthetic process; Enucleic acid binding; Elysine-tRNA ligase activity; Plysyl-tRNA aminoacylation; EATP binding; Plysine biosynthetic process	EC:6.1.1.6	IPRO04364; IPRO04365; IPRO06195; IPRO12340; IPRO16027; IPRO18149; IPRO18150; G3DSA:3.30.930.1 (GENE3D), SSF5568 (SUPERFAMILY)
	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](#)

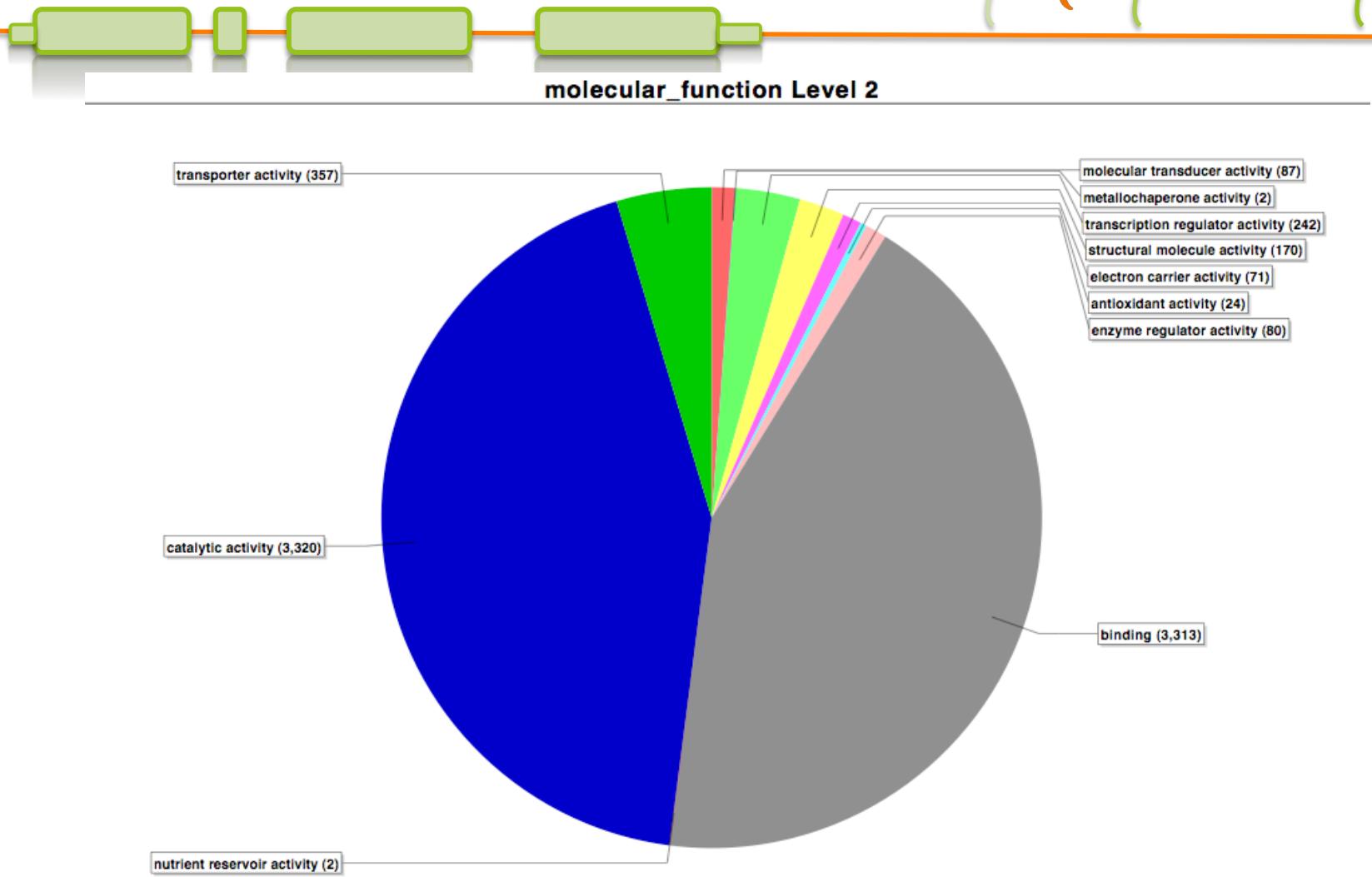
```

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

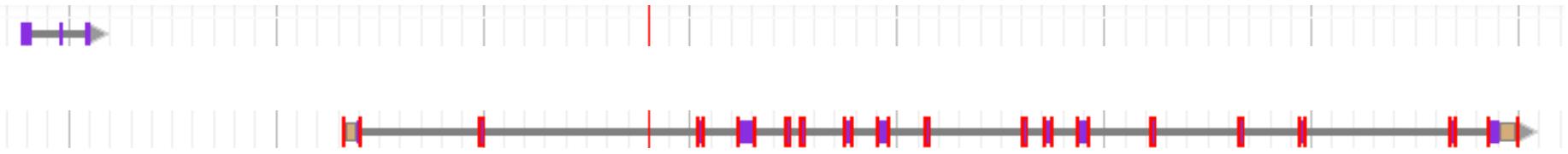




Quick view of synteny-based method

Liftovers are very useful for orthology determination

- Align two genomes (Satsuma)
- Transfer annotations between aligned regions (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

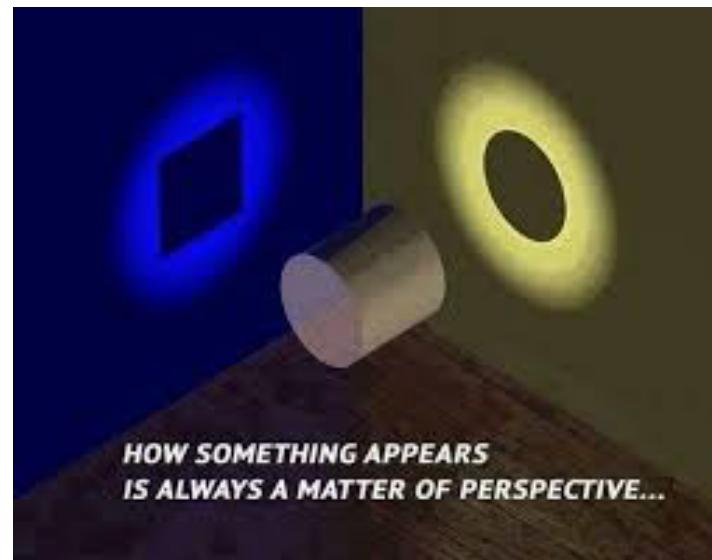
BUT

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 INFO

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
- Benzene 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
- Tropone, 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)



Jacques Dainat PhD

THE END

