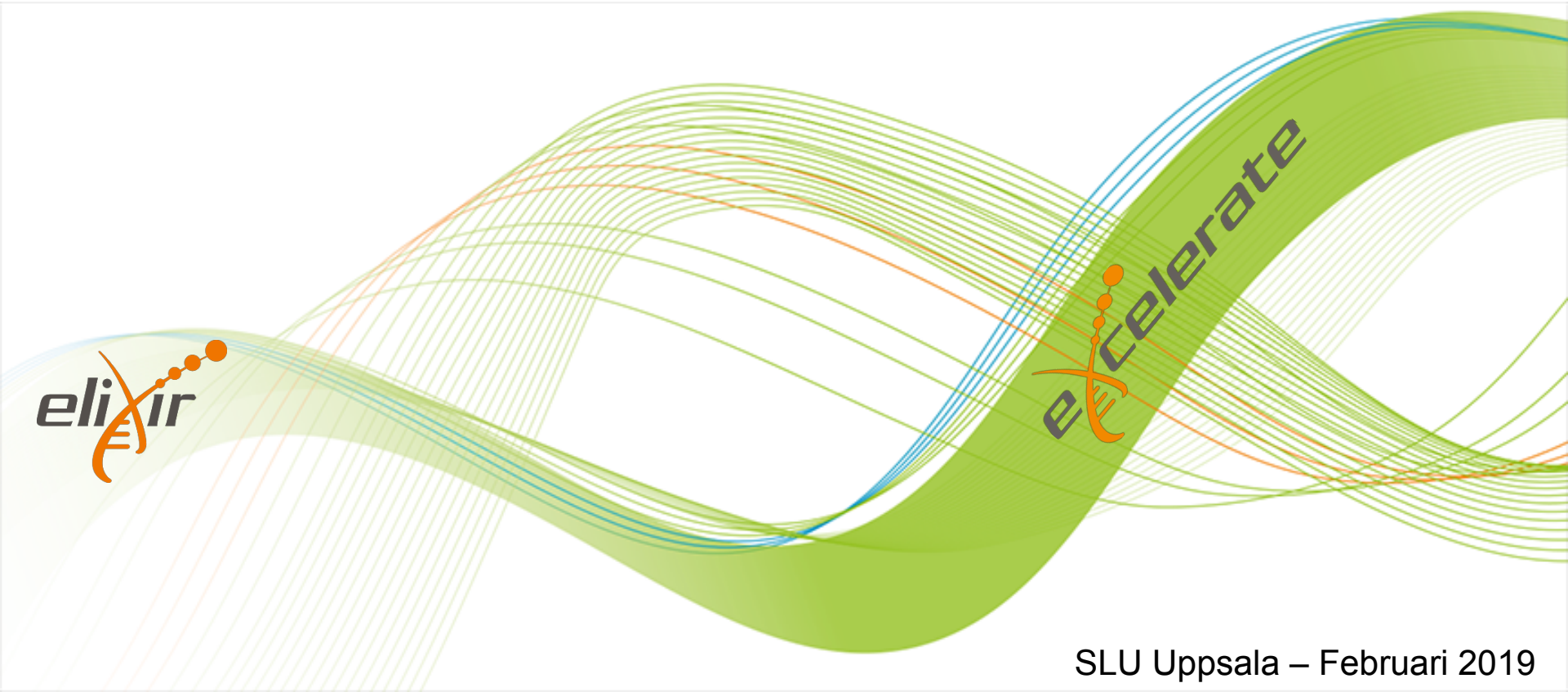
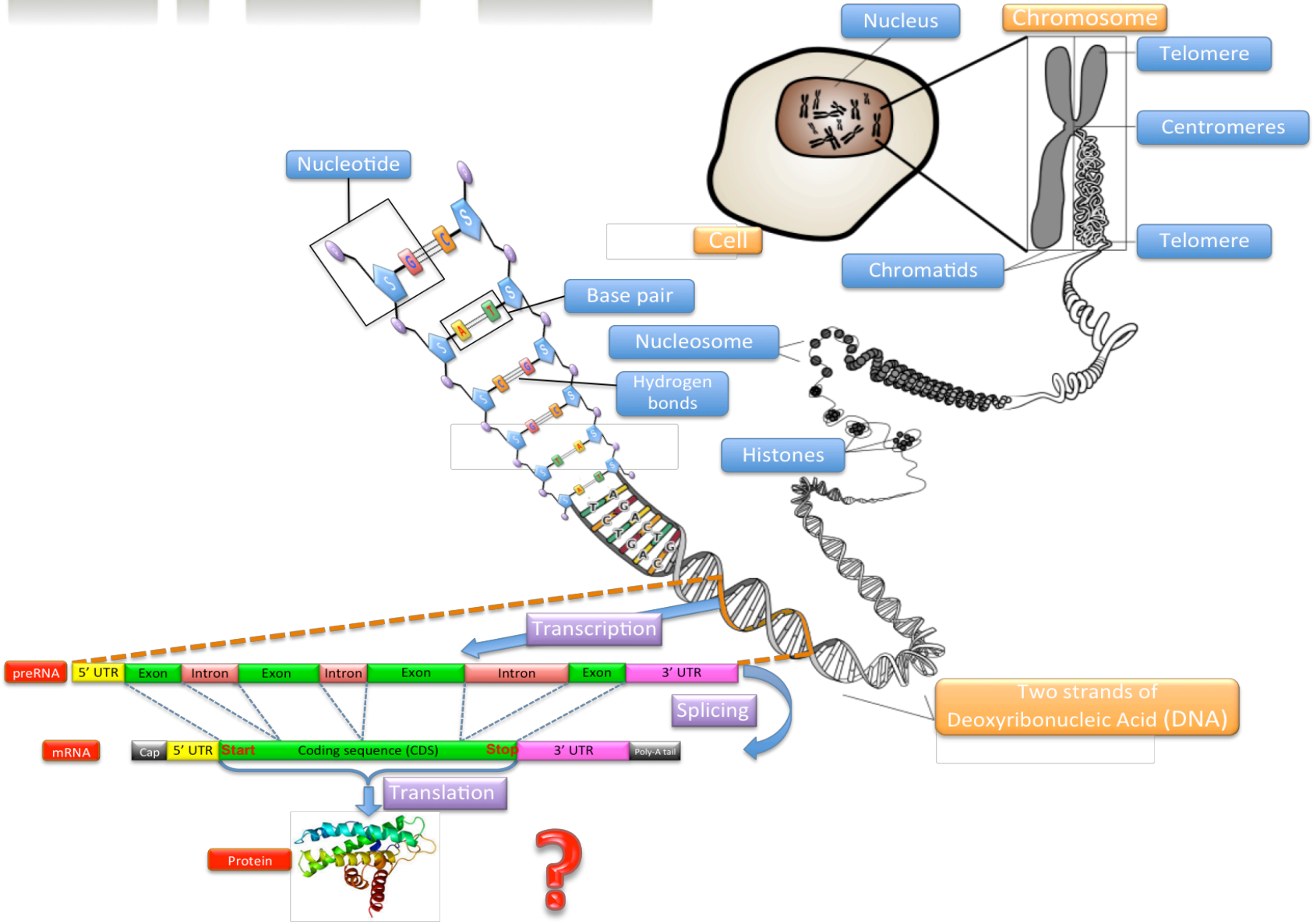
Jacques Dainat PhD

Functional annotation



SLU Uppsala – Februari 2019

Overview

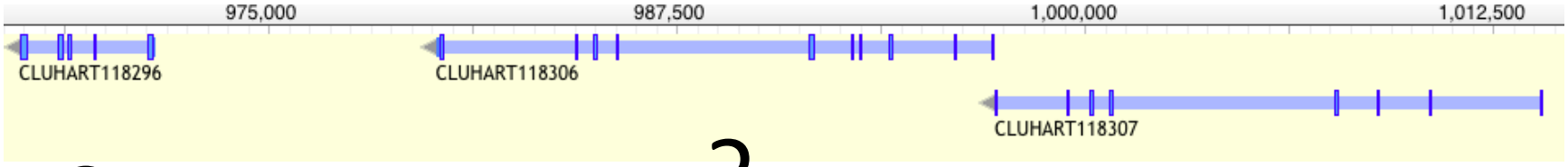

Nucleus
Chromosome
Telomere
Centromeres
Telomere
Chromatids
Cell
Nucleotide
Base pair
Nucleosome
Hydrogen bonds
Histones
Transcription
preRNA
5' UTR
Exon
Intron
Exon
Intron
Exon
Intron
Exon
3' UTR
Splicing
Two strands of Deoxyribonucleic Acid (DNA)
mRNA
Cap
5' UTR
Start
Coding sequence (CDS)
Stop
3' UTR
Poly-A tail
Translation
Protein
?

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.

Proteins vary in structure as well as function



?

Antibodies?

?

Energy?

?

Enzyme?

Contractile protein?

Transport Protein?

Hormone?

Storage Protein?

Structural Protein?

Functional annotation – HOW?

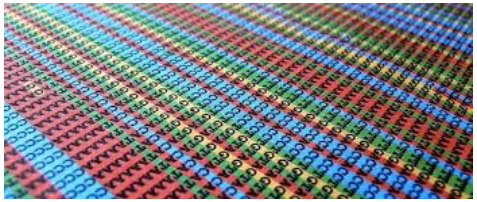
- Experimentally

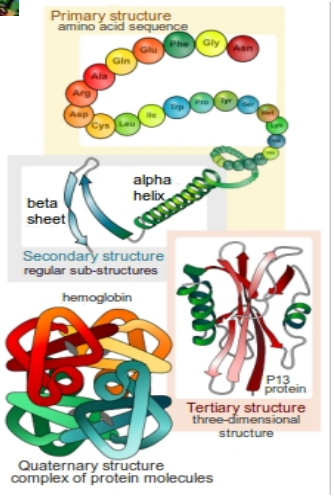
  => Mutants, knockout, etc.

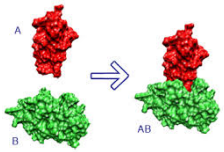
Accurate

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

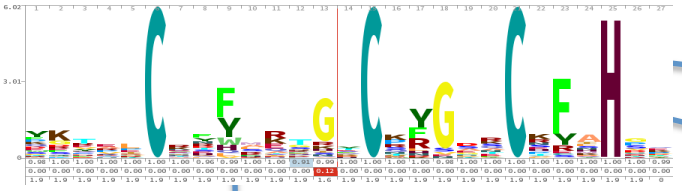
```
Q  GLMDTAFEHIKATGGLTTESNYPYKGEDATCNS-K
   GLM+ AFE+IK +GG+TTES YPY+  + TC++ +
S  GLMENAFEYIKHSGGITTESAYPYRAANGTCDAVR
```

- Based on Motif

  ⇒Proscan, MEME, QuasiMotiFinder

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 / COG
  - Based on synteny

    ⇒Whole genome alignment (lastZ)

    (NBIS) Satsuma + kraken + custom script
  - Based on phylogeny

    ⇒ Quite complicated at large scale

Methods - **Structure-based**

- 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

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).

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

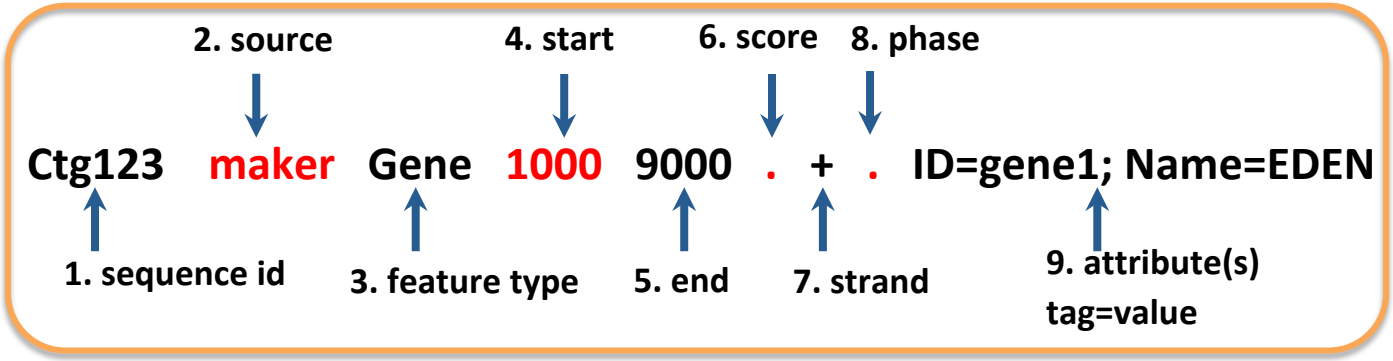
Functional annotation – HOW?

Get sequences

Functional annotation – HOW?

- Genome is in fasta format.

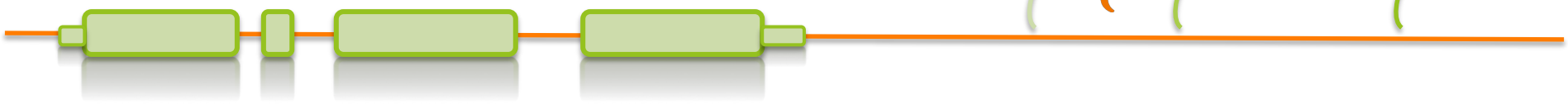
- Annotation is often in GFF-format. This format contains in general only coordinates, but sometimes it can include the sequence as well.

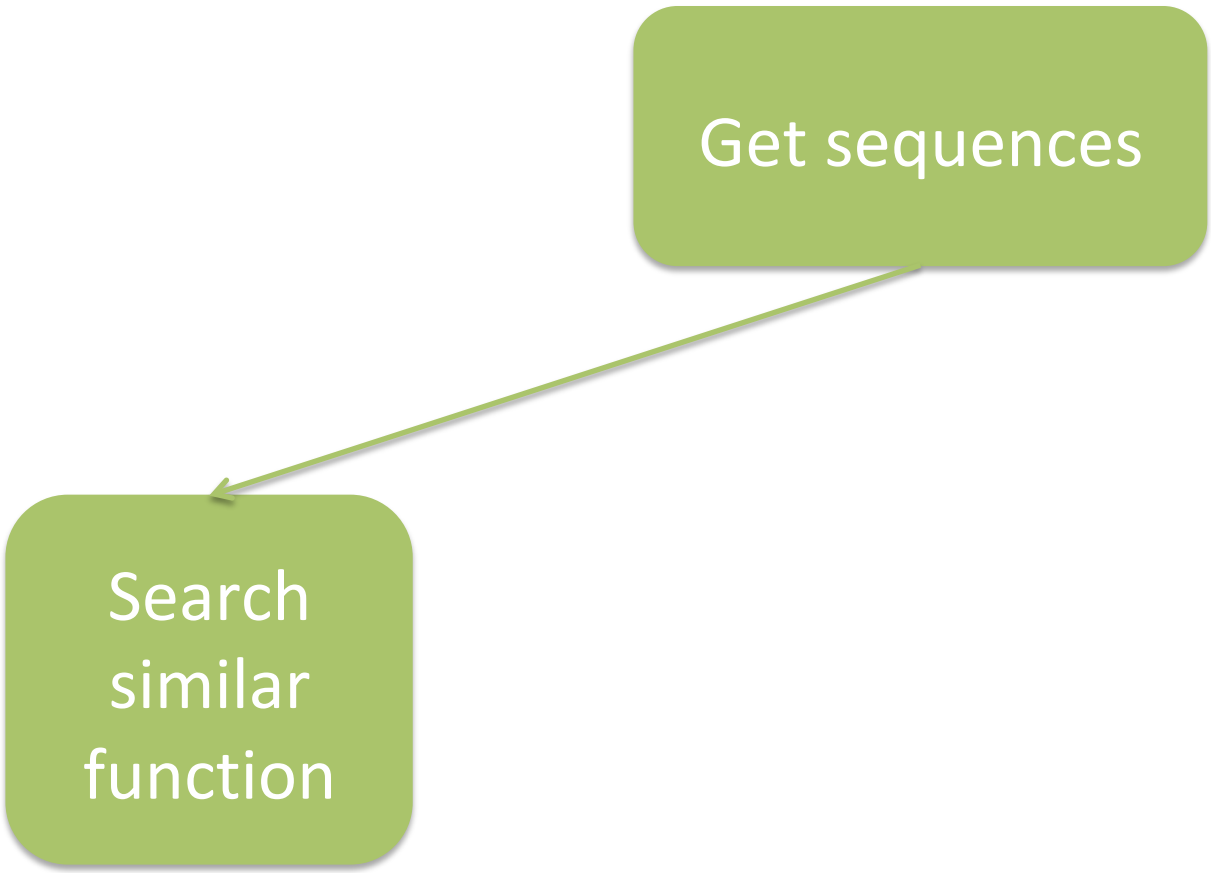


- You can use the GFF-file together with the genome-file to extract the gene sequences.

- 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?





Blast-based approach

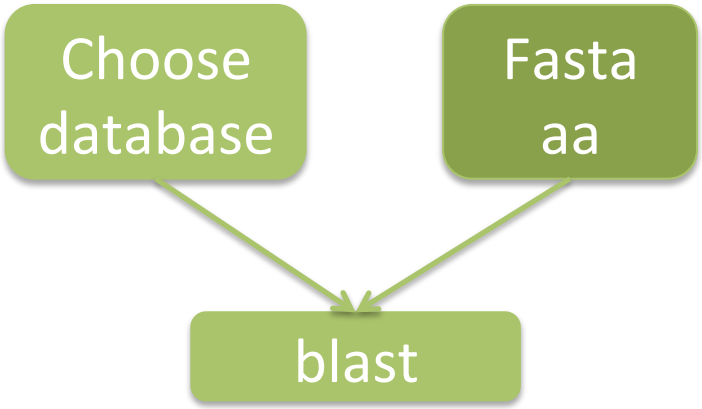
Annotate the sequences functionally using Blast

Choose database

Uniprot	Swissprot
exhaustive	reliable

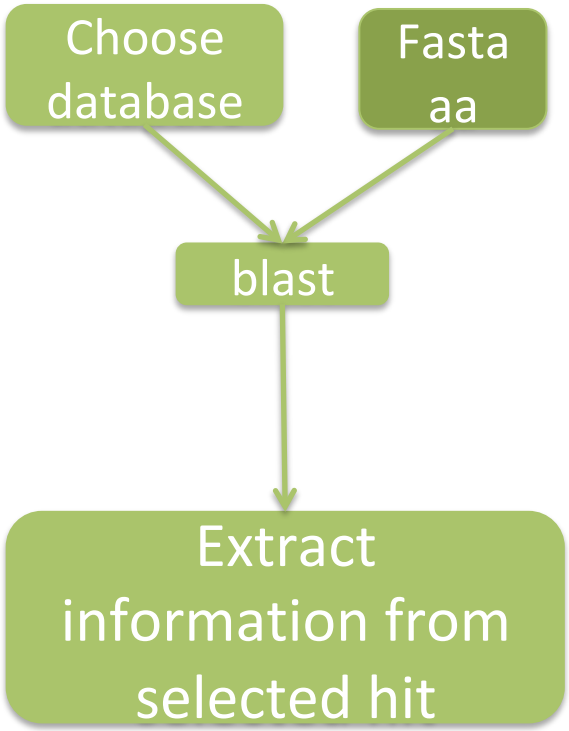
Blast-based approach

Annotate the sequences functionally using Blast



Minimum Threshold

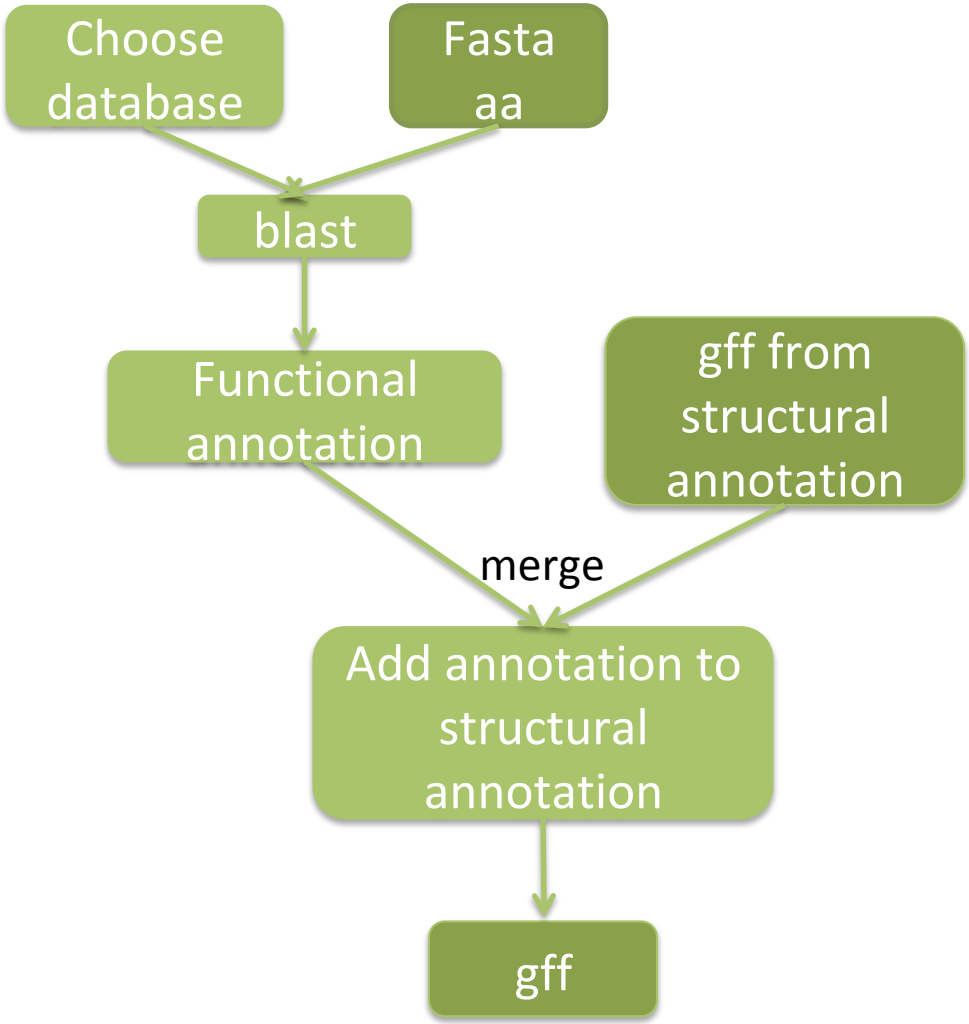
Annotate the sequences functionally using Blast



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)

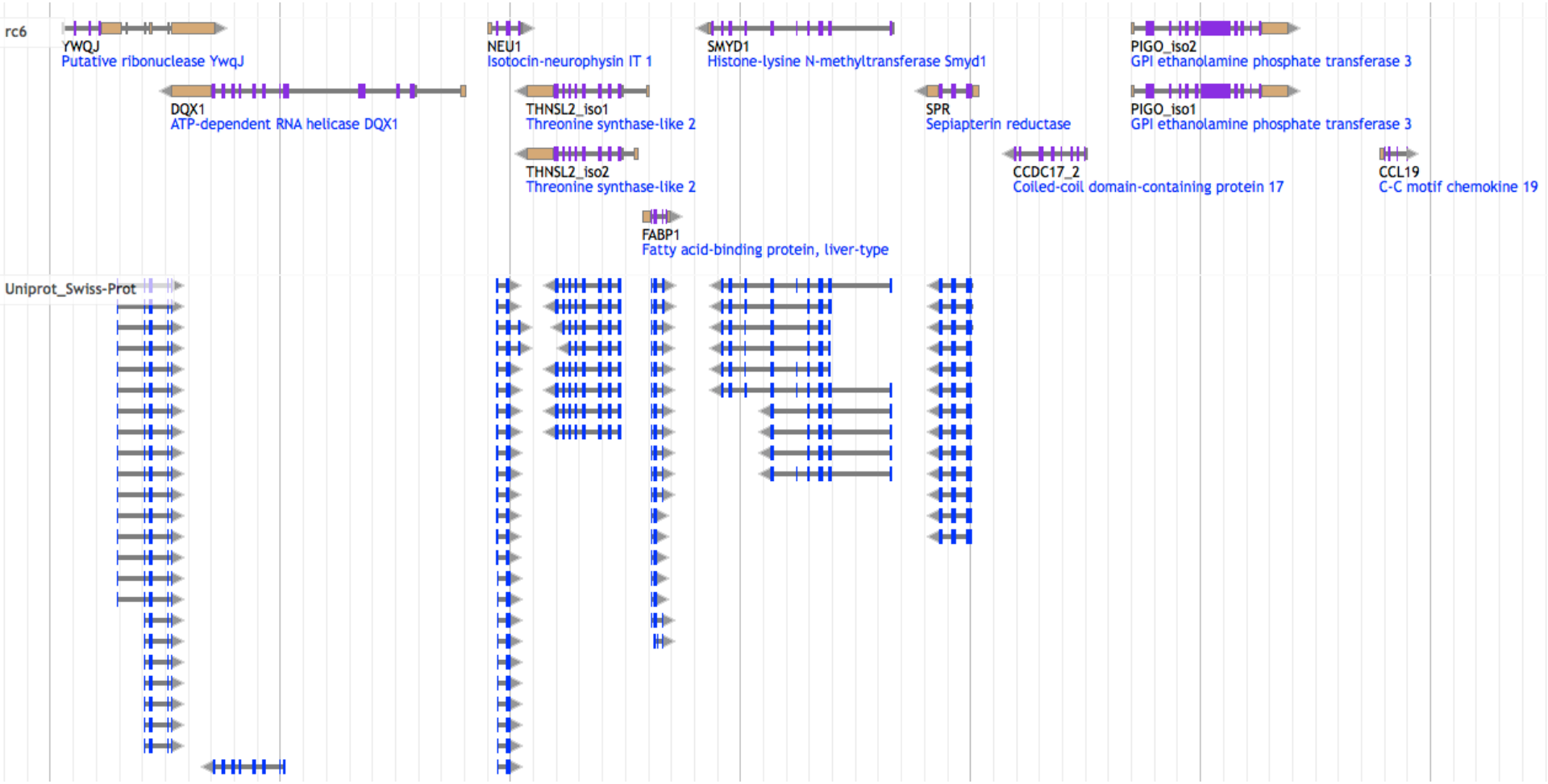
Limits

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

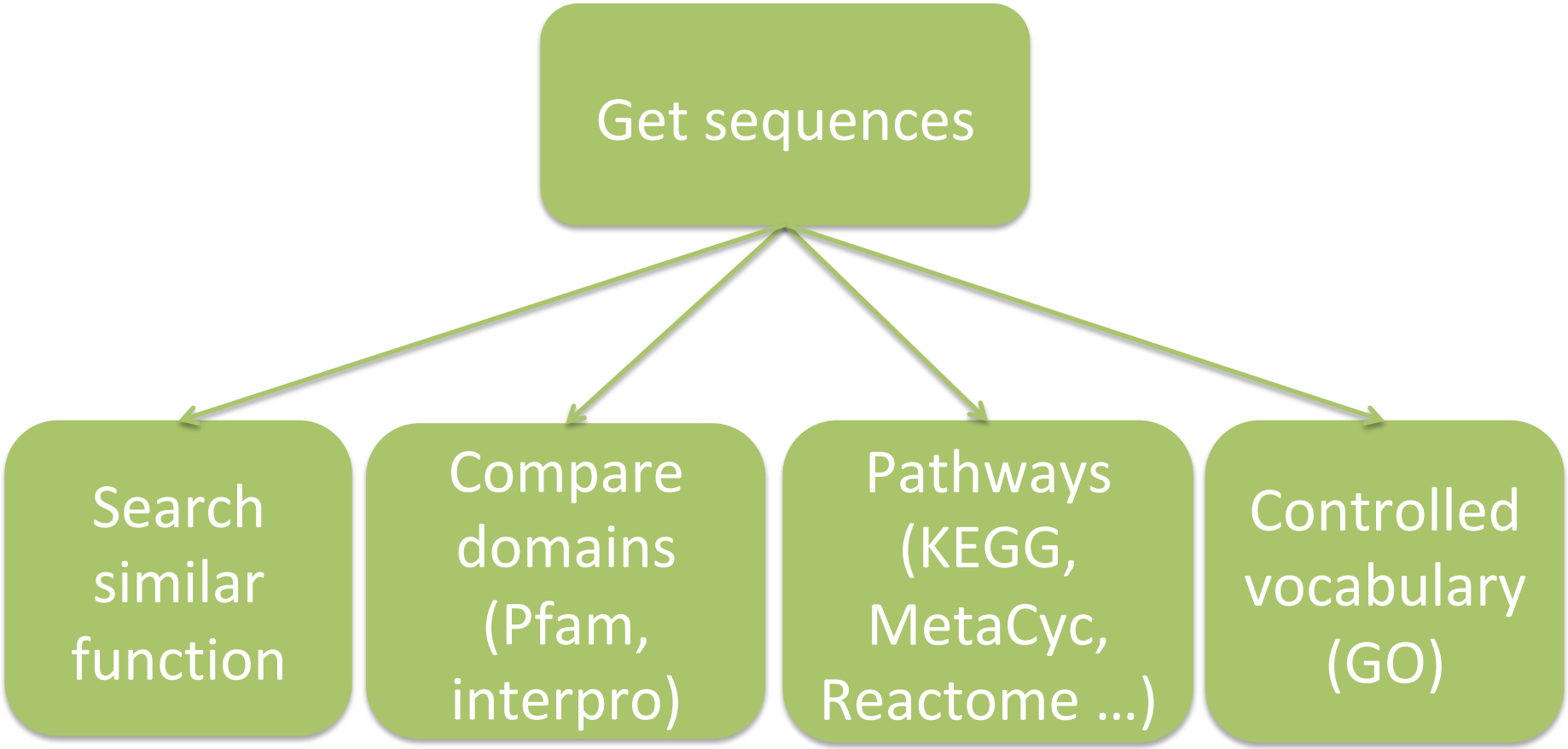
Blast-based annotation are tightly dependent to the quality of the structural annotation

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

Blast-based approach : result



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:

More than 60 000 terms

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

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

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

molecular activities of gene products

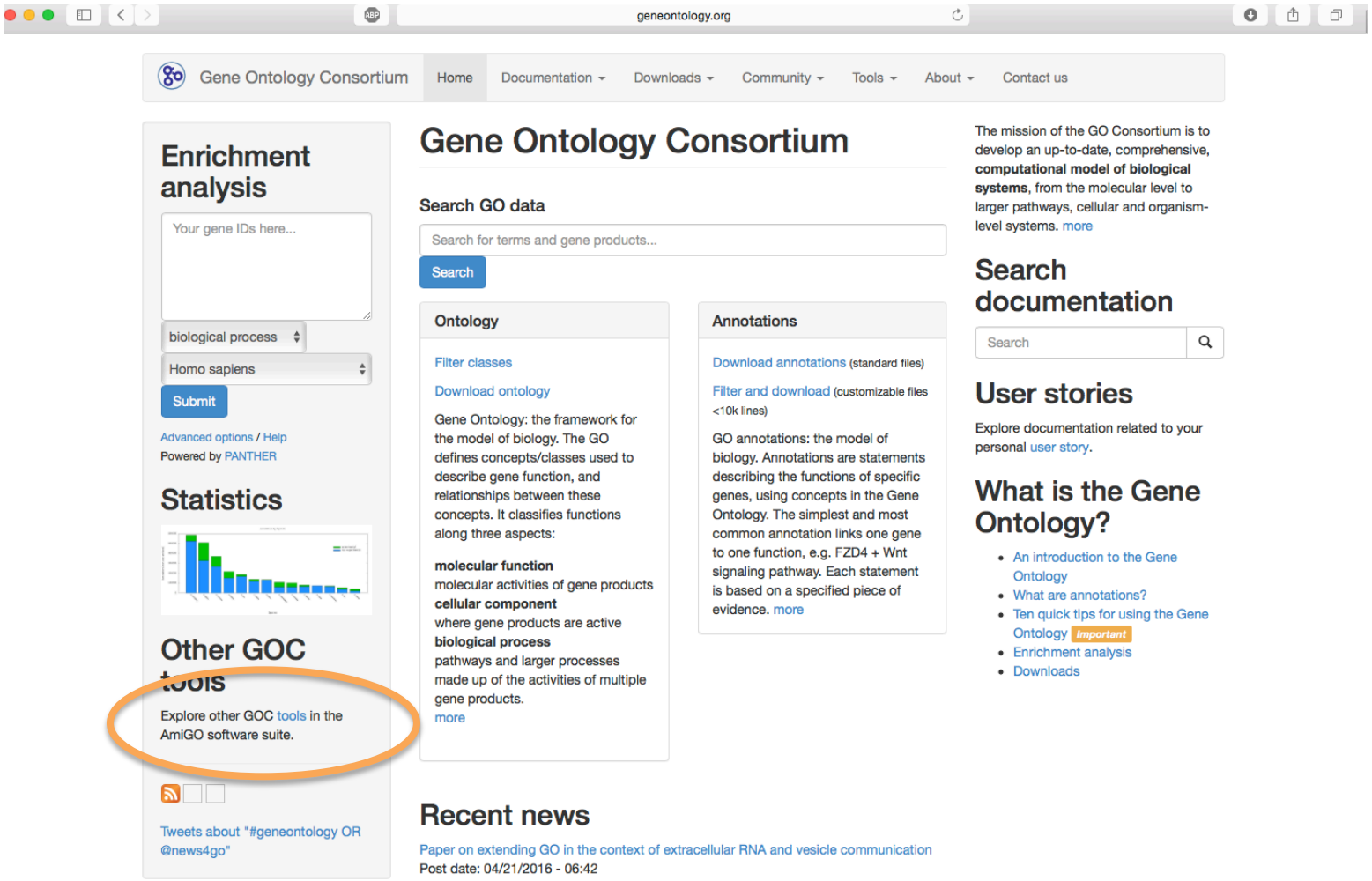
Cellular Component

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

where gene products are active

Gene Ontology

http://www.geneontology.org/



Tools

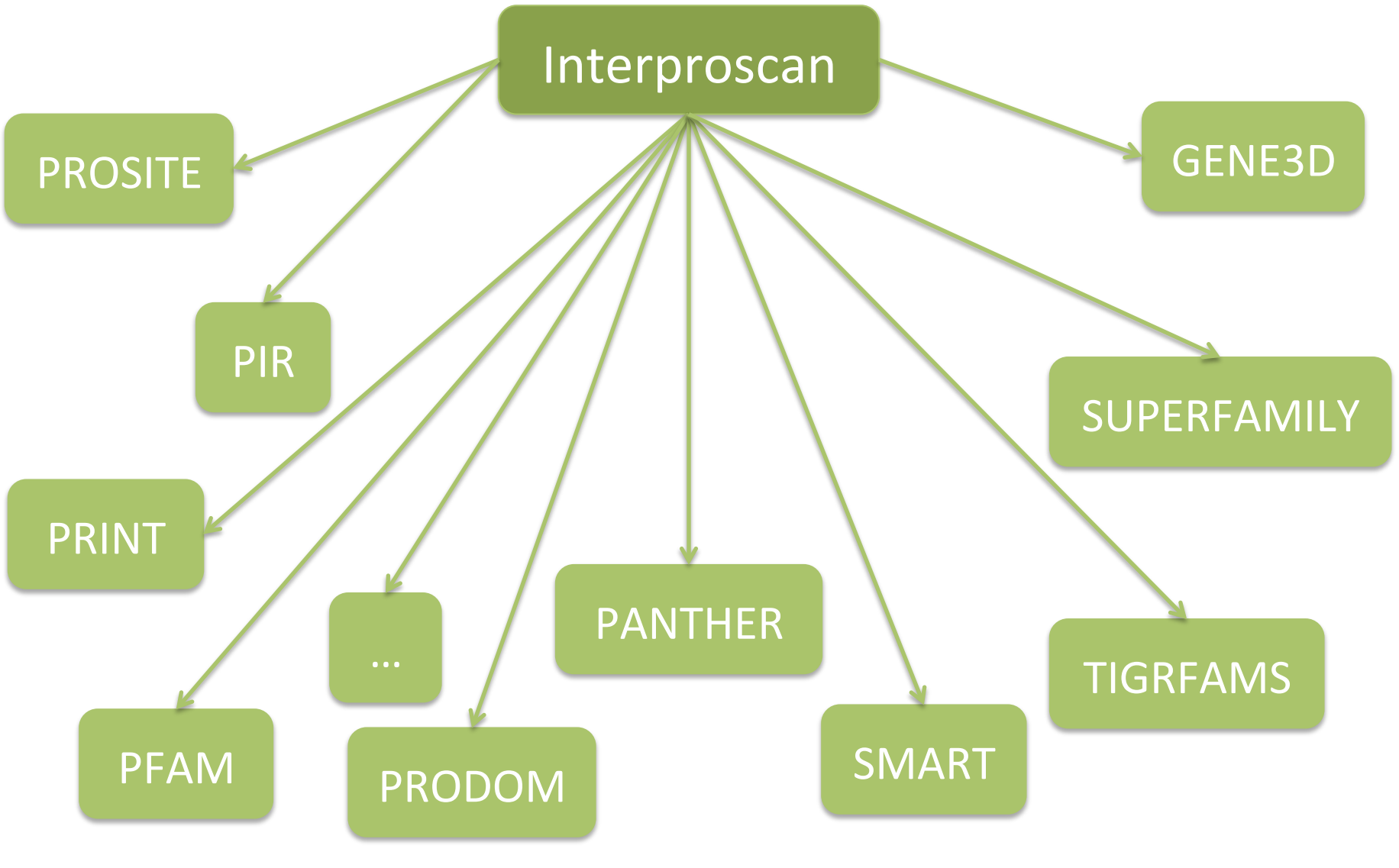
Tool	Approach	Comment
Trinotate	Best blast hit + protein domain identification (HMMER/ PFAM) + protein signal peptide and transmembrane domain prediction (signalP/tmHMM), and leveraging various annotation databases (eggNOG/GO/Kegg databases).	Not automated
Annocript	Best blast hit	Collects the best-hit and related annotations (proteins, domains, GO terms, Enzymes, pathways, short)
Annot8r	Best blast hit**s**	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 hit**s	**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.”

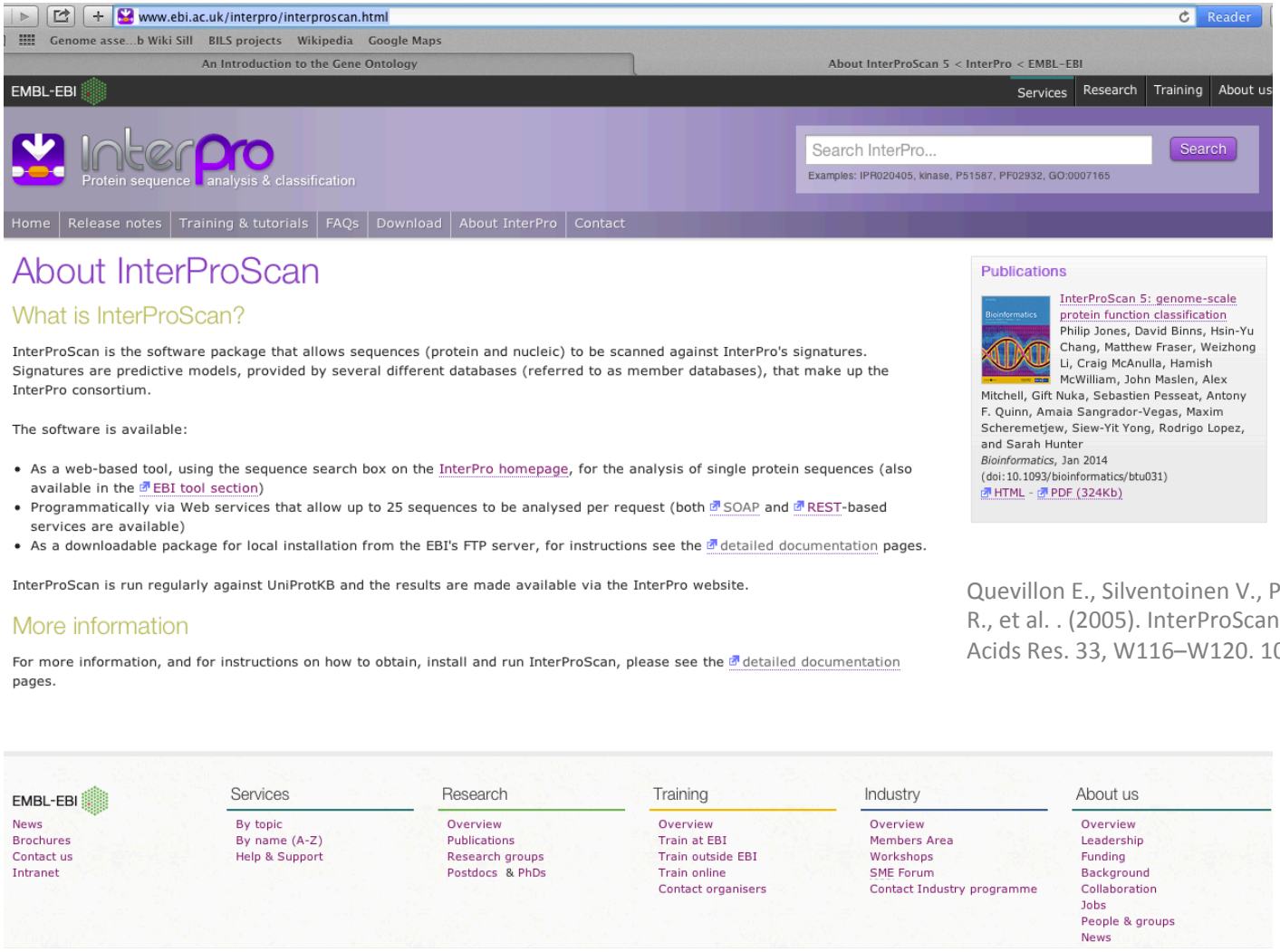
https://www.ebi.ac.uk/interpro/about.html

Interproscan



Interproscan

- Annotate the sequences functionally using Interproscan



Jones,P.etal.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

Interproscan

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:

- F Family (19869)
- D Domain (8868)
- R Repeat (282)
- S 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:

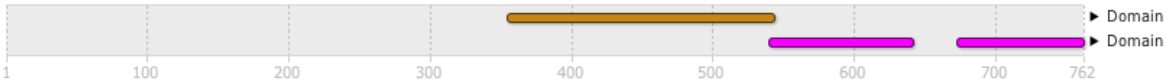
- KEGG
- MetaCyc
- Reactome
- UniPathway

Interproscan results

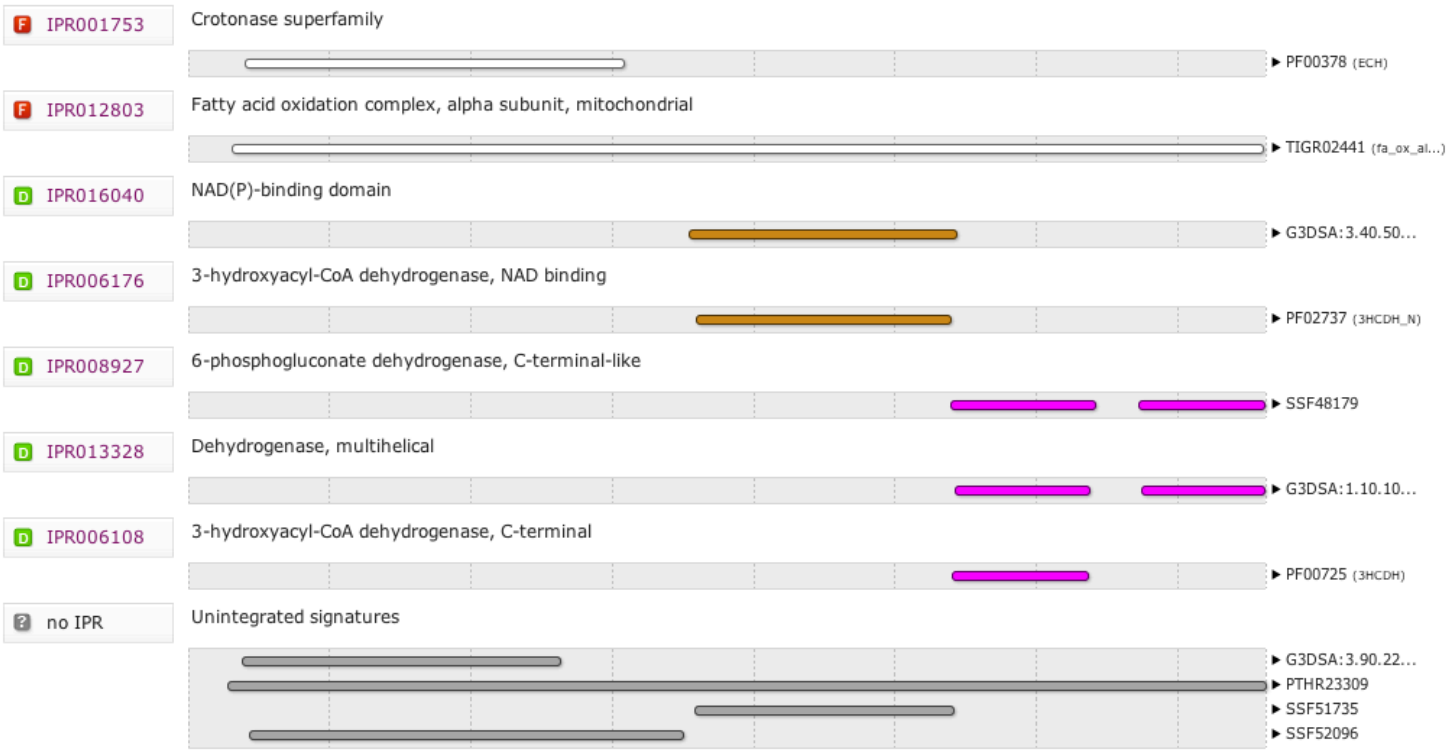
Protein family membership

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

Domains and repeats



Detailed signature matches



Interproscan results

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

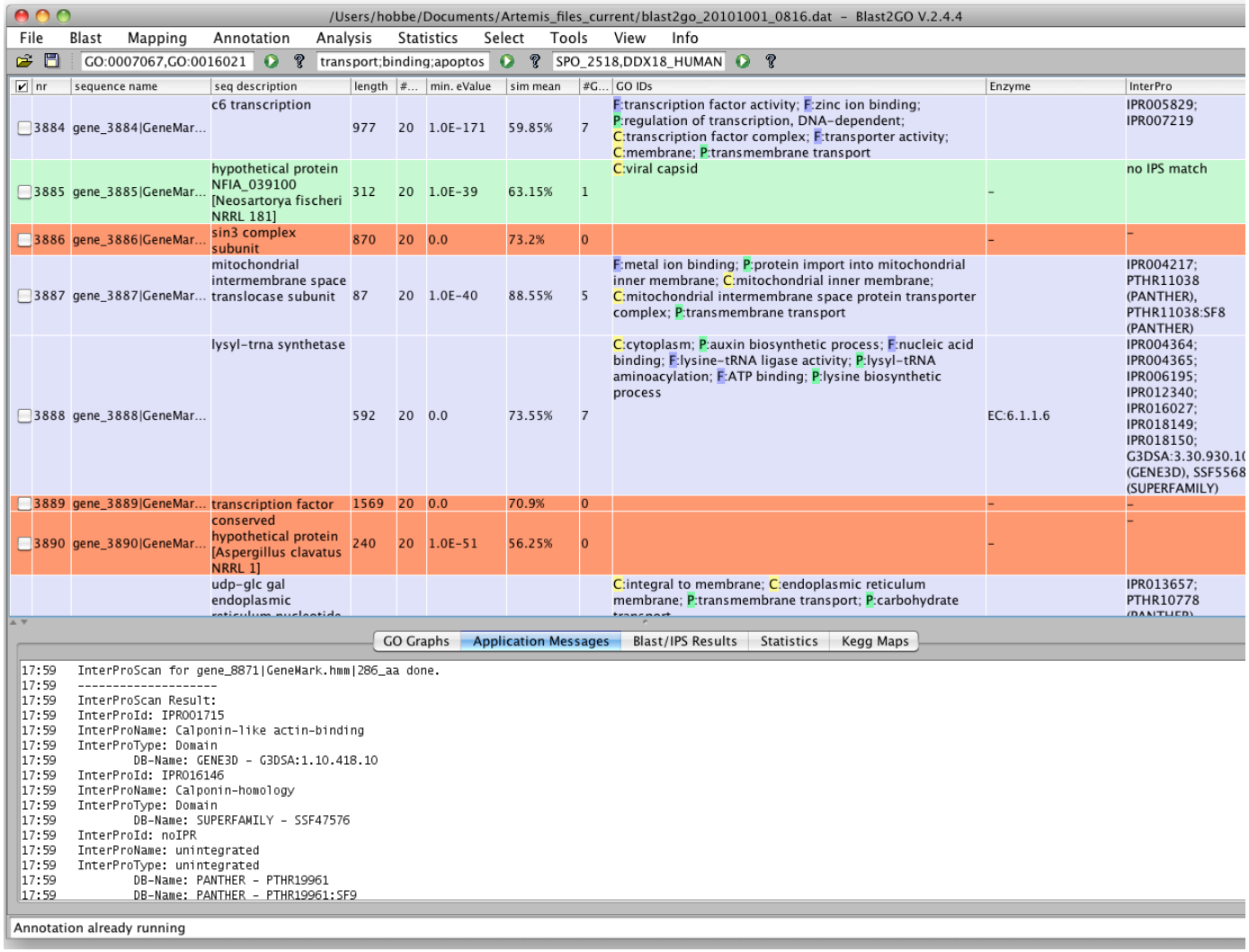
Blast2GO

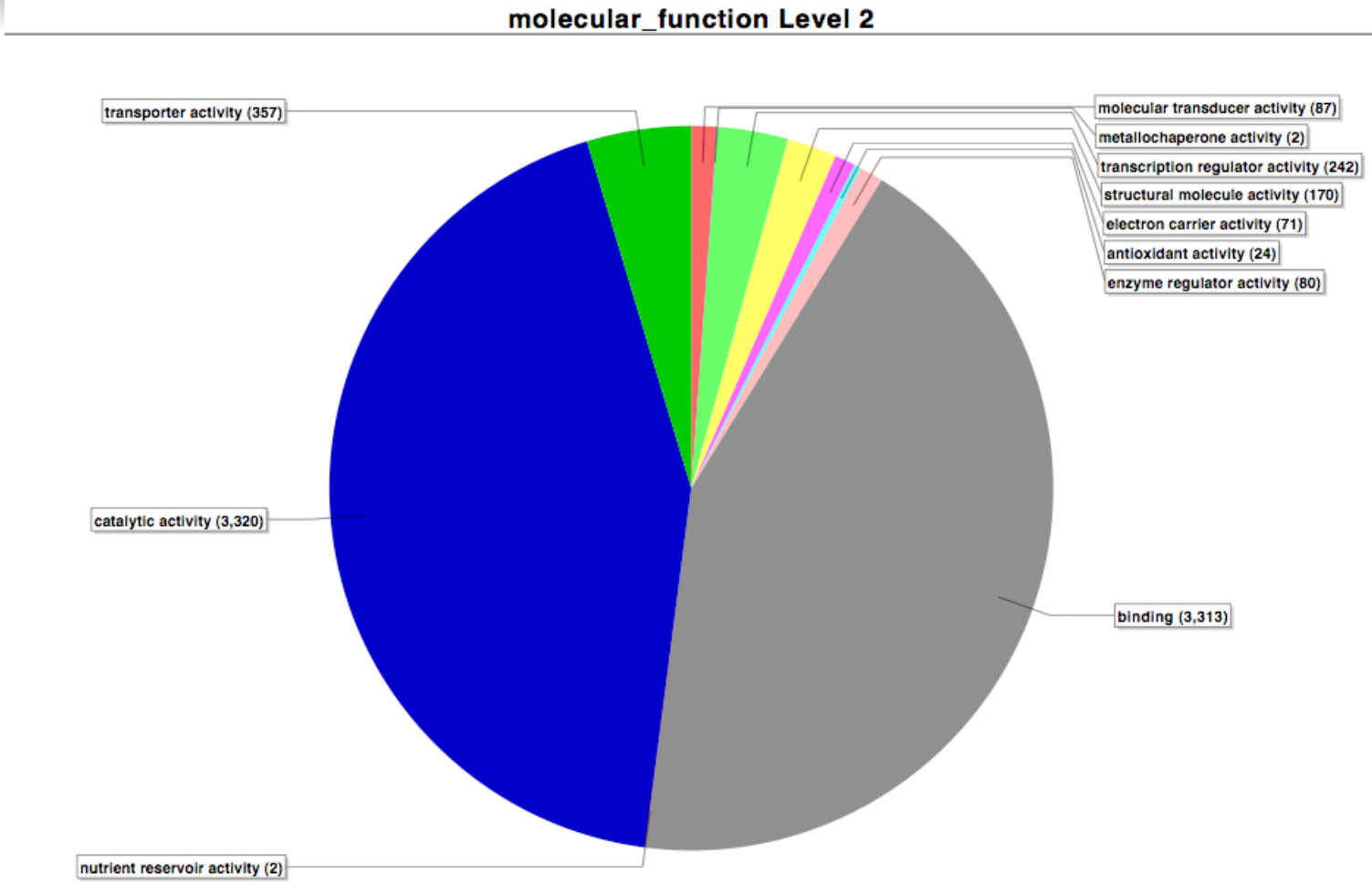
Another way : use the (mostly) commercial alternative

- 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

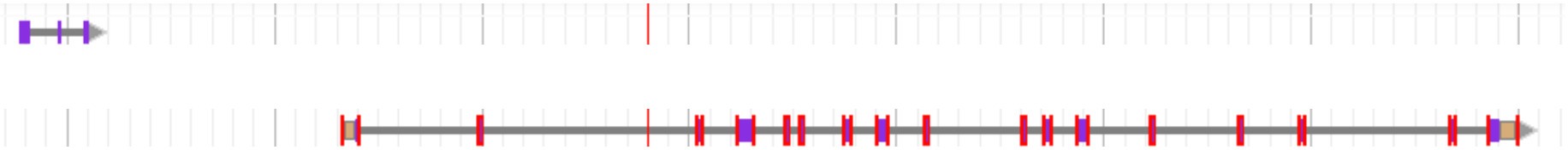

/Users/hobbe/Documents/Artemis_files_current/blast2go_20101001_0816.dat - Blast2GO V.2.4.4
File Blast Mapping Annotation Analysis Statistics Select Tools View Info
GO:0007067,GO:0016021
transport;binding;apoptos
SPO_2518,DDX18_HUMAN
nr | sequence name | seq description | length | #... | min. eValue | sim mean | #G... | GO IDs | Enzyme | InterPro
3884 | gene_3884|GeneMar... | c6 transcription | 977 | 20 | 1.0E-171 | 59.85% | 7 | F:transcription factor activity; F:zinc ion binding; P:regulation of transcription, DNA-dependent; C:transcription factor complex; F:transporter activity; C:membrane; P:transmembrane transport | | IPR005829; IPR007219
3885 | gene_3885|GeneMar... | hypothetical protein NFIA_039100 [Neosartorya fischeri NRRL 181] | 312 | 20 | 1.0E-39 | 63.15% | 1 | C:viral capsid | - | no IPS match
3886 | gene_3886|GeneMar... | sin3 complex subunit | 870 | 20 | 0.0 | 73.2% | 0 | | - | -
3887 | gene_3887|GeneMar... | mitochondrial intermembrane space translocase subunit | 87 | 20 | 1.0E-40 | 88.55% | 5 | F:metal ion binding; P:protein import into mitochondrial inner membrane; C:mitochondrial inner membrane; C:mitochondrial intermembrane space protein transporter complex; P:transmembrane transport | | IPR004217; PTHR11038 (PANTHER), PTHR11038:SF8 (PANTHER)
3888 | gene_3888|GeneMar... | lysyl-trna synthetase | 592 | 20 | 0.0 | 73.55% | 7 | C:cytoplasm; P:auxin biosynthetic process; F:nucleic acid binding; F:lysine-tRNA ligase activity; P:lysyl-tRNA aminoacylation; F:ATP binding; P:lysine biosynthetic process | EC:6.1.1.6 | IPR004364; IPR004365; IPR006195; IPR012340; IPR016027; IPR018149; IPR018150; G3DSA:3.30.930.1((GENE3D), SSF5568 (SUPERFAMILY)
3889 | gene_3889|GeneMar... | transcription factor | 1569 | 20 | 0.0 | 70.9% | 0 | | - | -
3890 | gene_3890|GeneMar... | conserved hypothetical protein [Aspergillus clavatus NRRL 1] | 240 | 20 | 1.0E-51 | 56.25% | 0 | | - | -
udp-glc gal endoplasmic | | | | | | | C:integral to membrane; C:endoplasmic reticulum membrane; P:transmembrane transport; P:carbohydrate | | IPR013657; PTHR10778
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: IPR001715
17:59 InterProName: Calponin-like actin-binding
17:59 InterProType: Domain
17:59 DB-Name: GENE3D - G3DSA:1.10.418.10
17:59 InterProId: IPR016146
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


molecular_function Level 2
transporter activity (357)
molecular transducer activity (87)
metallochaperone activity (2)
transcription regulator activity (242)
structural molecule activity (170)
electron carrier activity (71)
antioxidant activity (24)
enzyme regulator activity (80)
catalytic activity (3,320)
binding (3,313)
nutrient reservoir activity (2)

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

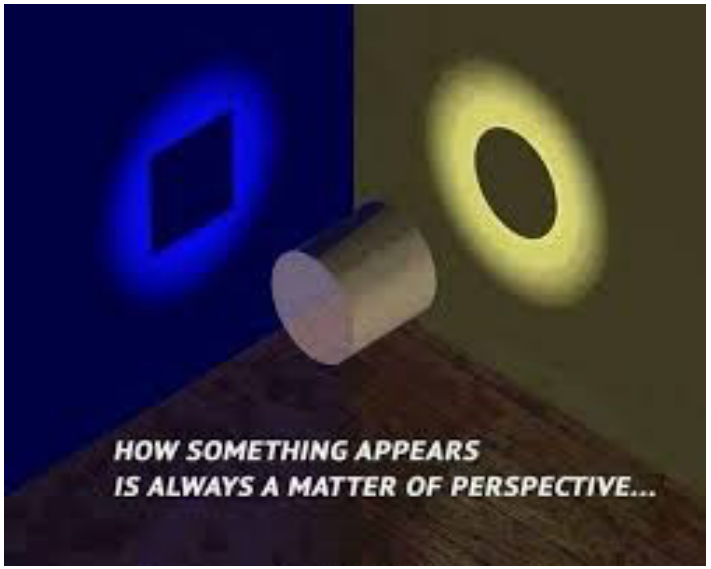
BUT

gene has no function alone

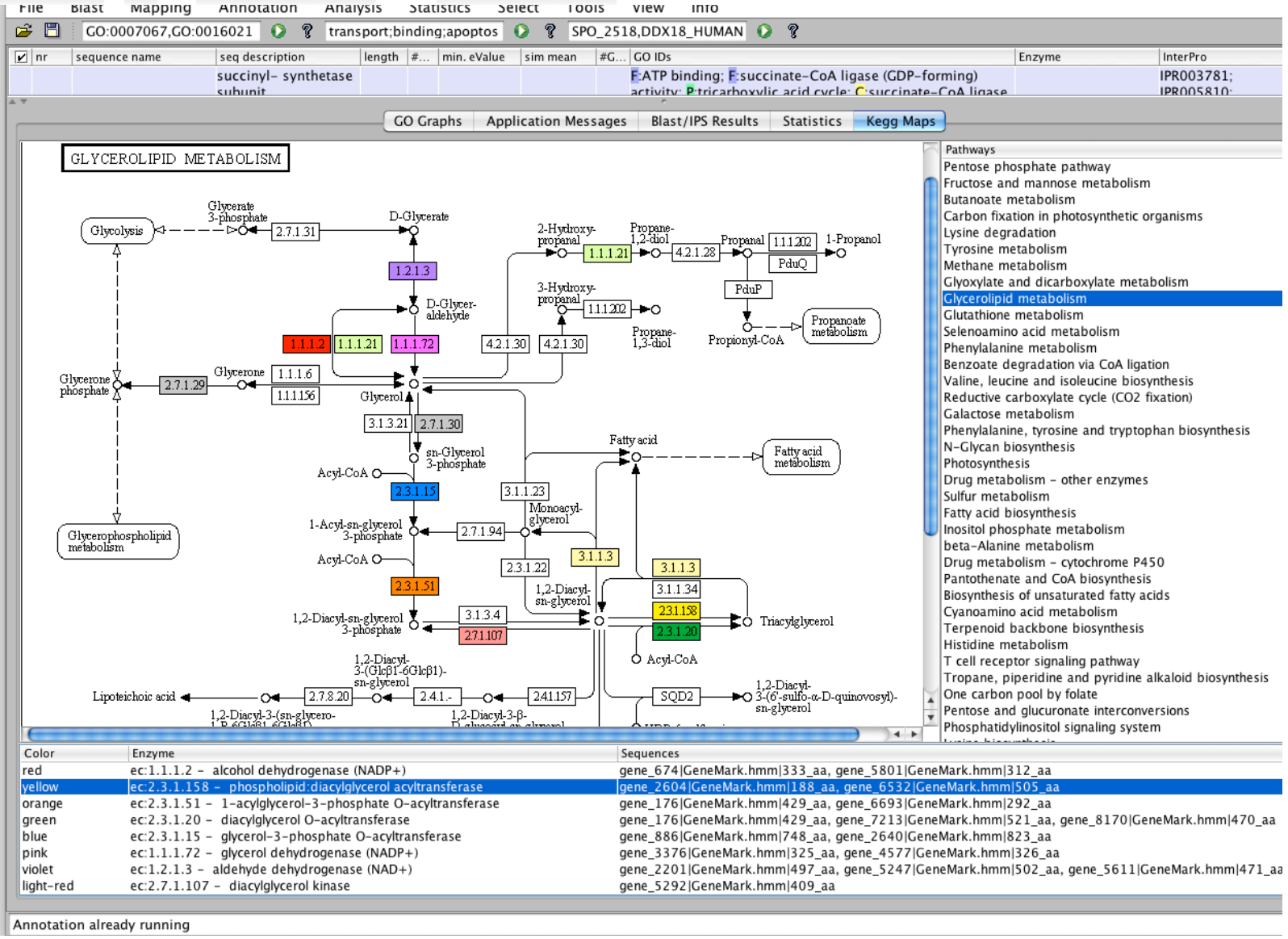
⇒ Pathways / regulatory networks explain how genes interact so what they are doing!

E.g. databases for pathway :

- KEGG
- MetaCyc
- Reactome
- UniPathway



KEGG-mapping



Conclusion

- Functional annotation found

  /!\ Transmission of error from databases !

  Experimental check is good !

- Hypothetical protein / Uncharacterized protein

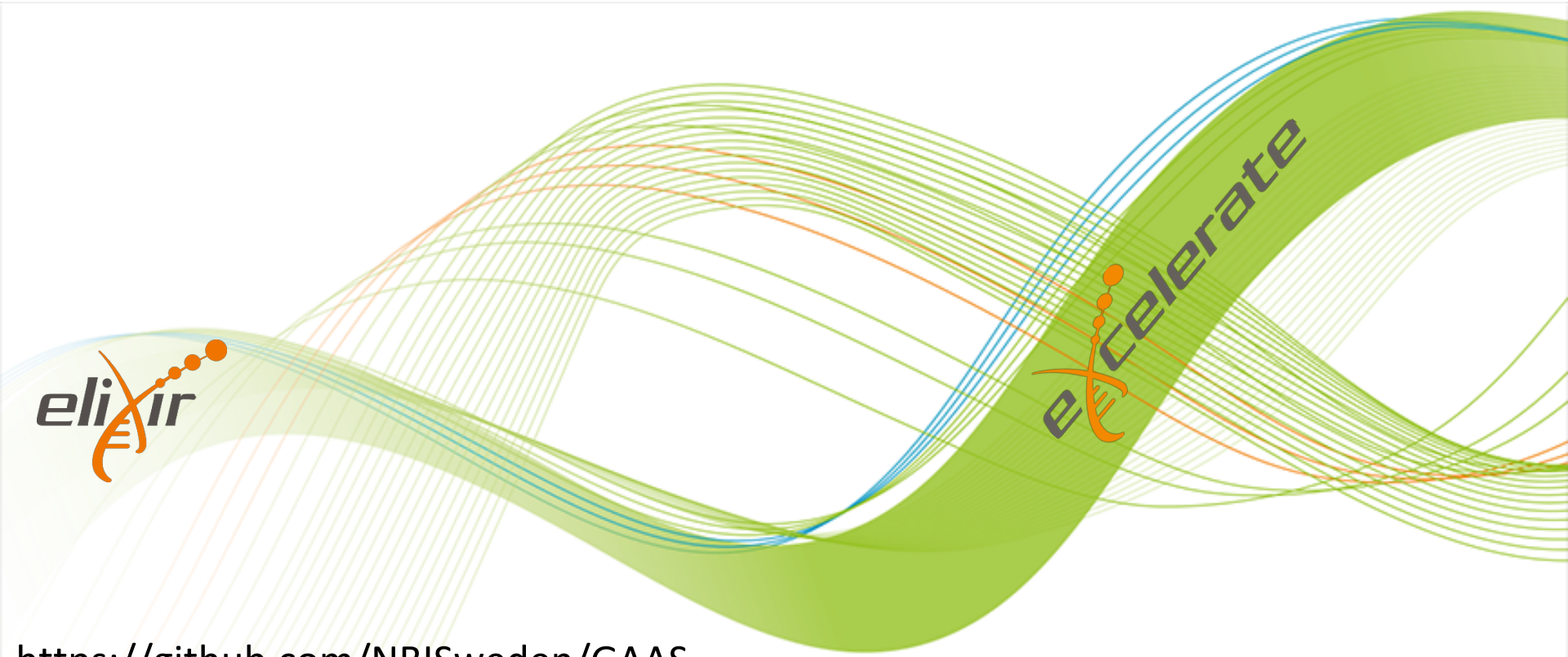
  => depends largely on conventional experiments.

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



https://github.com/NBISweden/GAAS