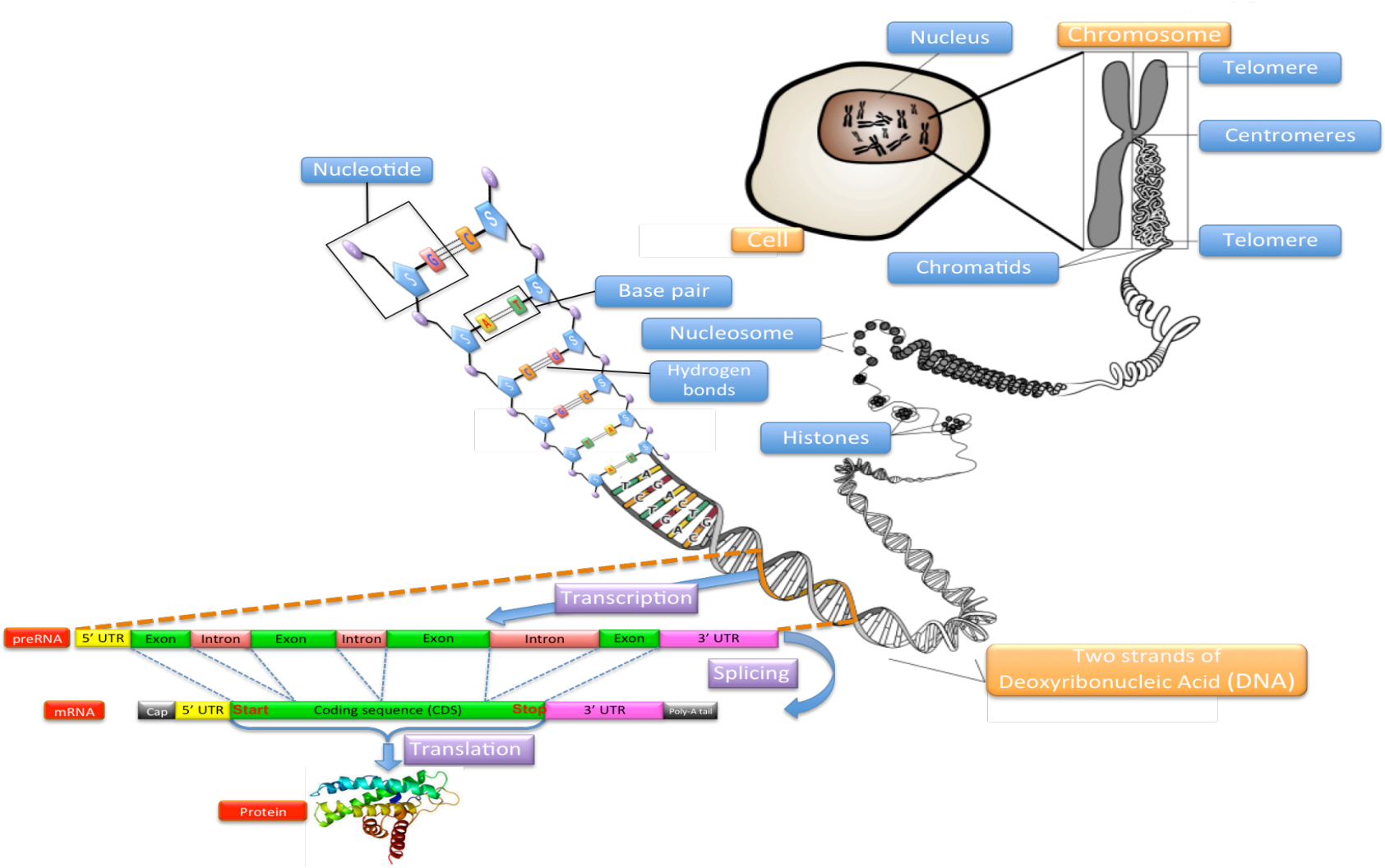
Functional annotation



Right, now we have our genes, but what do they do ?

?

?

?

Insulin receptor?

Vesicle-trafficking protein?

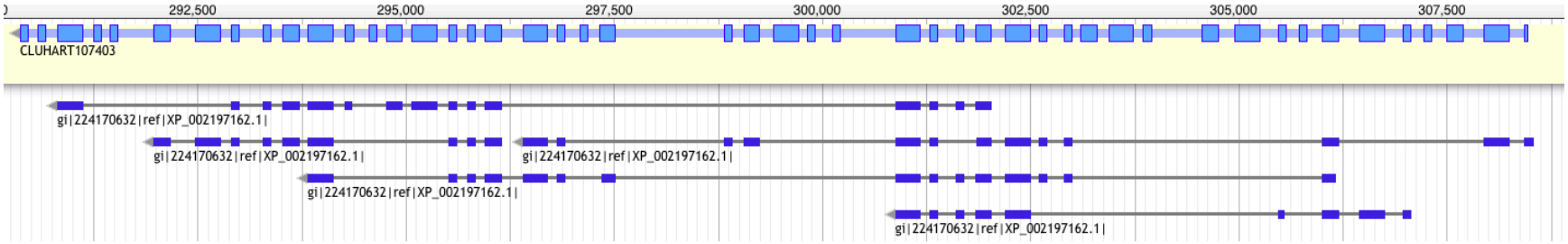
Alcohol dehydrogenase?

Aquaporin?

Transcription factor

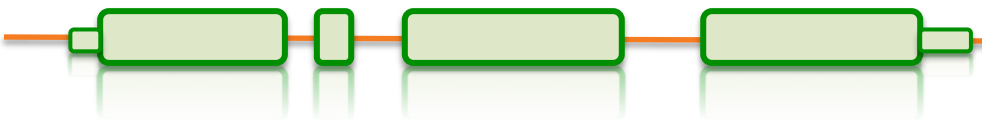
MAP kinase kinase kinase?

But we have used proteins in our annotation!



It is actually kind of complex…

… and most of pipelines do not do this for you.



First you need the sequences

- Extract sequences from the browser (Webapollo)
- GFF3 => fasta : Use gffread (in Cufflinks package)
- Fasta available (Biomart, FTP, output of annotation tools)
- If CDS=> translate in AA : Use gffread (in Cufflinks package)

Methods

- Based on blast
  - Best blast hit
  - Clustering
- Based on synteny

  ⇒ Whole genome alignment (lastZ)

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

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 !

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.		

Starting point for downstream analysis

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

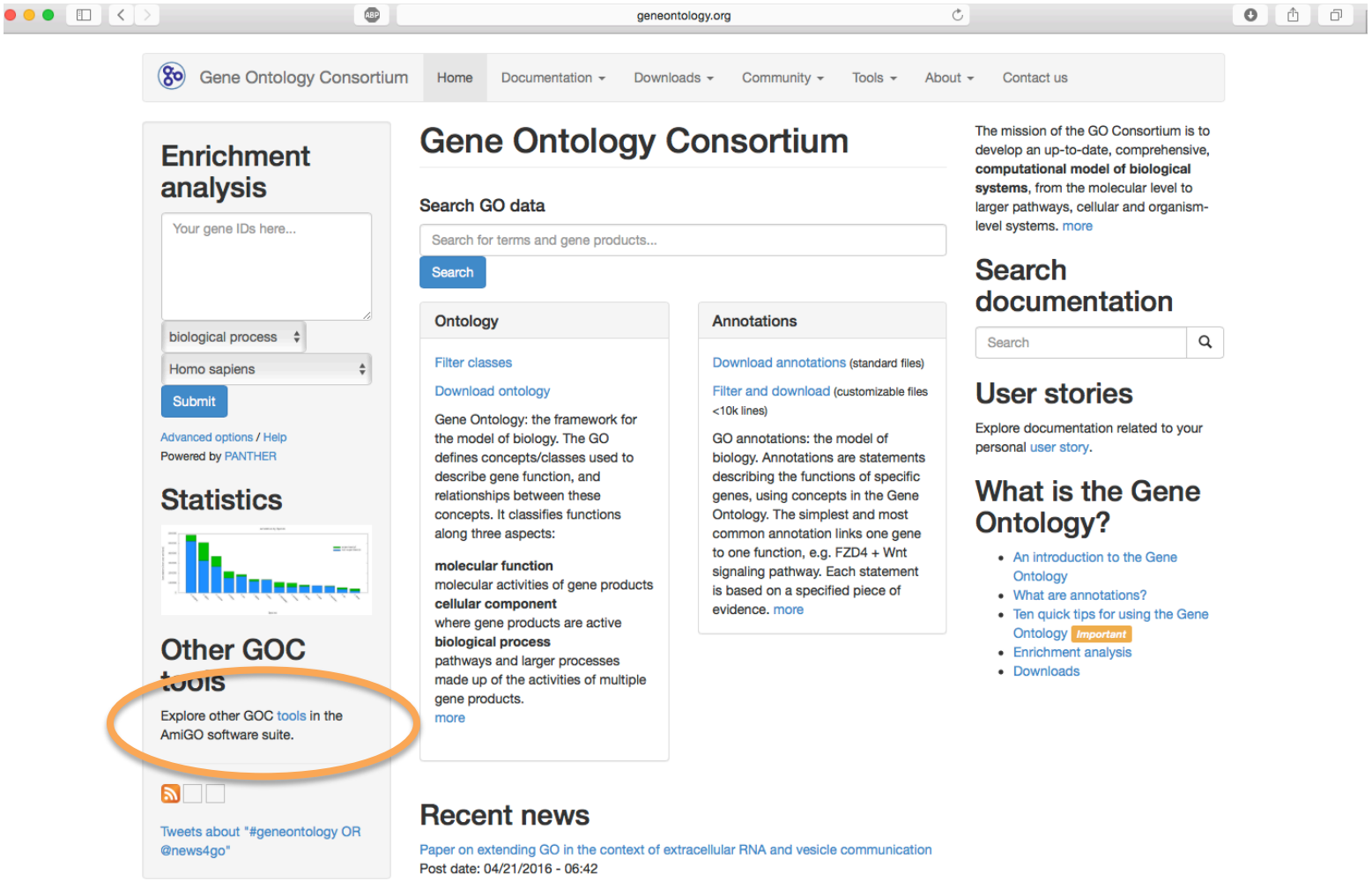
Cellular Component

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

More than 60 000 terms

Gene Ontology

Starting point of downstream analysis



There is an equivalence table between GO and Interpro

GO annotation is given with an Evidence Code:

- IDA: inferred by direct assay
- ISS: infered by sequence similarity
- IEA: electronic annotation
- ....

Search applications have two main modalities

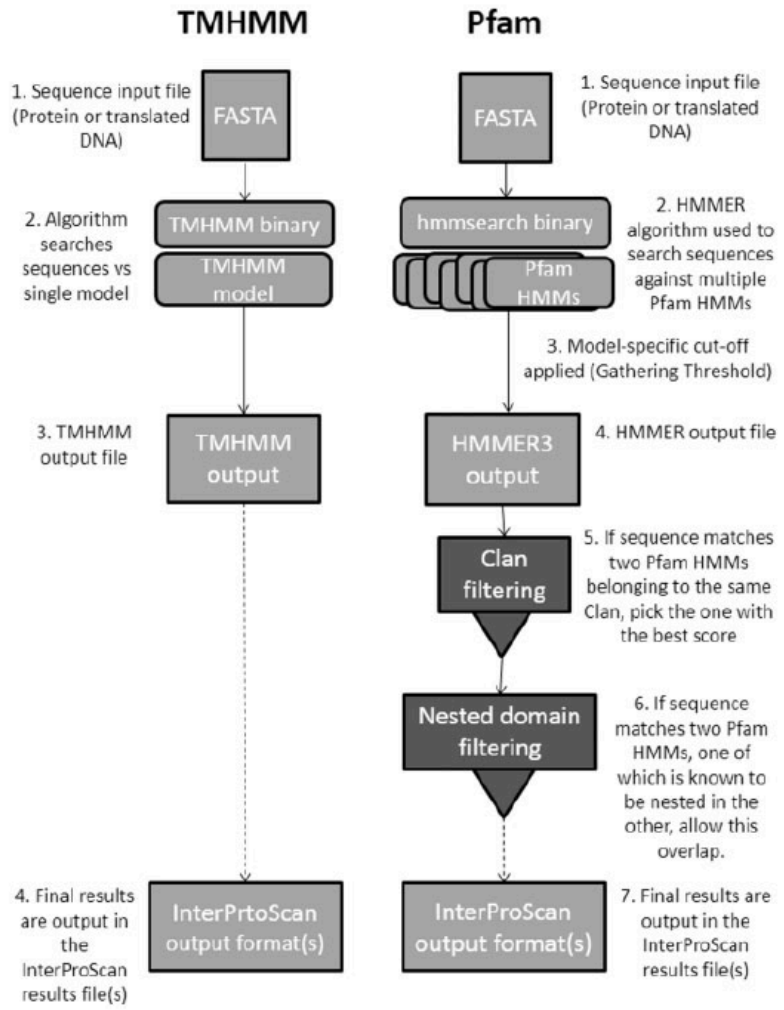
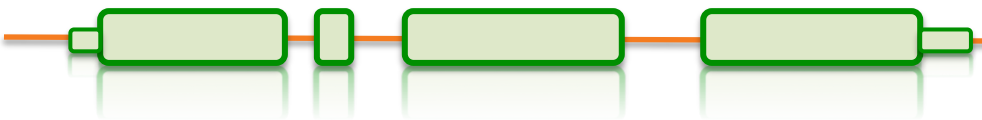


Fig. 1. Comparison of the processing steps used by two different member database applications, TMHMM and Pfam

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

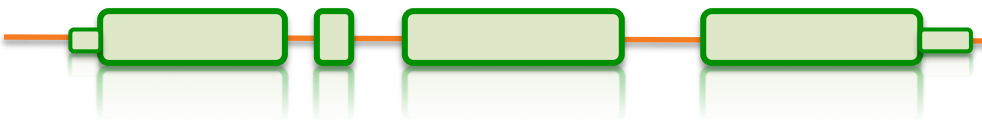
Blast-based approach



Annotate the sequences functionally using Blast

- Blast the protein-sequences from your maker-run to a protein blast-database (e.g., uniprot) using blastp from the Blast+ package
- Use Annie to extract best hits from blast-hit list and the corresponding description from uniprot-headers
- Add this information to the annotation.gff using custom-script

Blast-based approach



Strengths

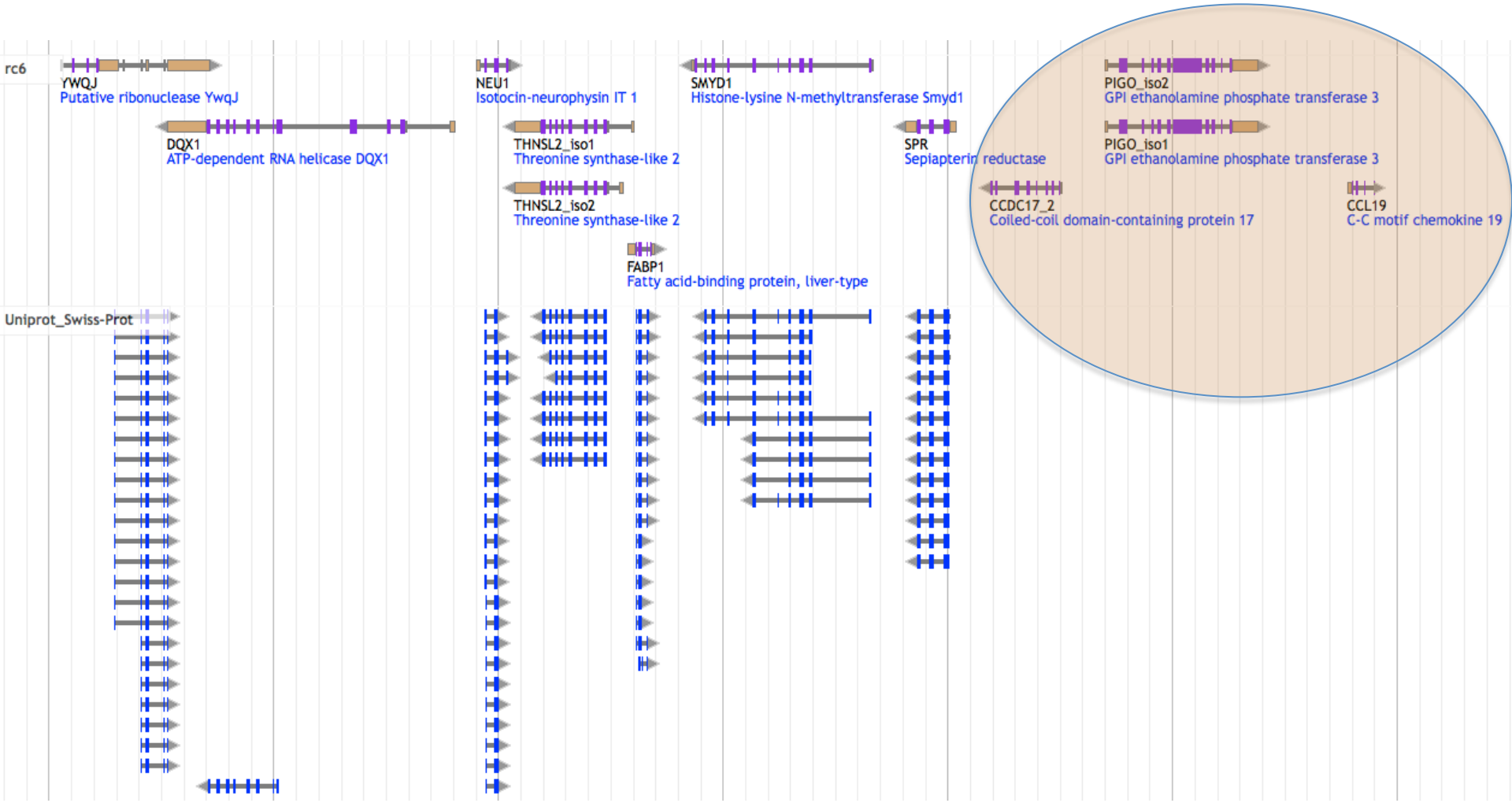
- Fairly fast and easy
- Allow gene naming

Limits

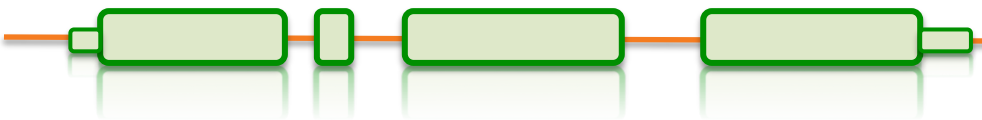
- Orthology not certain - best blast-hit does not equal orthologous!
- Bias due to well conserved domains
- Best Hit (use as template) is not necessary the best annotated sequence to use => Could apply a prioritization rule (Human first, then mouse, etc).

Blast-based approach

Aligned proteins ≠ Blast hits



Blast-based approach

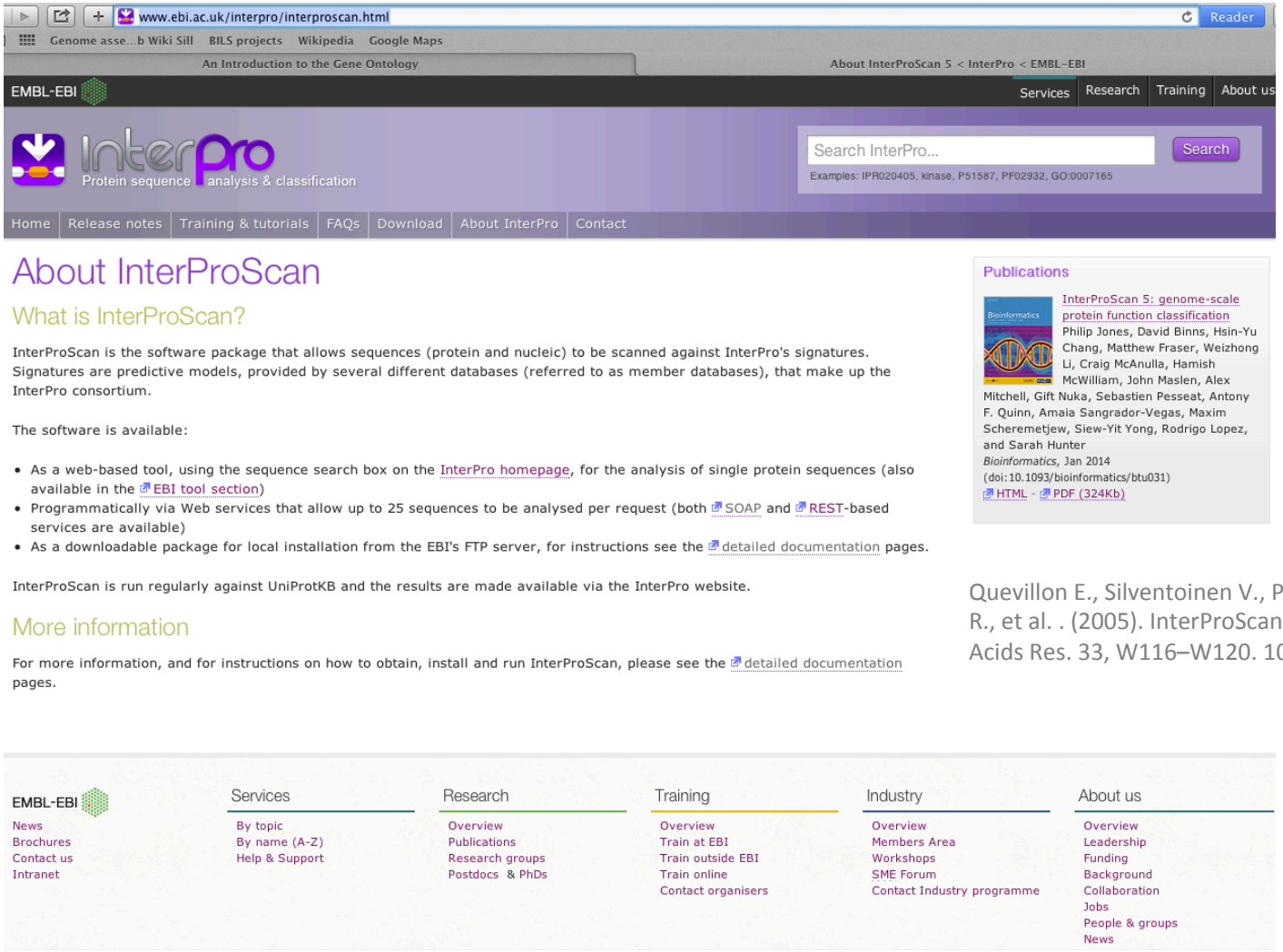


Blast-based annotation are tightly dependent to the quality of the annotation

- Gene Fusion
- Gene split
- Gene Partial (Well conserved domain)

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 57.0

InterPro protein matches are now calculated for all UniProtKB and UniParc proteins. The following statistics are for all UniProtKB proteins.
InterPro release 57.0 contains 29175 entries (last entry: IPR033481), representing:

- F Family (19597)
- D Domain (8393)
- R Repeat (284)
- S Sites
  - Active site (129)
  - Binding site (75)
  - Conserved site (680)
  - PTM (17)

InterPro cites 47947 publications in PubMed.

Member database information

Signature database	Version	Signatures*	Integrated signatures**
CATH-Gene3D	3.5.0	2626	1725
HAMAP	201511.02	2045	2037
PANTHER	10.0	95118	5179
PIRSF	3.01	3285	3223
PRINTS	42.0	2106	2002
PROSITE patterns	20.119	1309	1291
PROSITE profiles	20.119	1136	1108
Pfam	29.0	16295	15700
ProDom	2006.1	1894	1130
SMART	7.1	1312	1265
SUPERFAMILY	1.75	2019	1416
TIGRFAMs	15.0	4488	4454

* 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

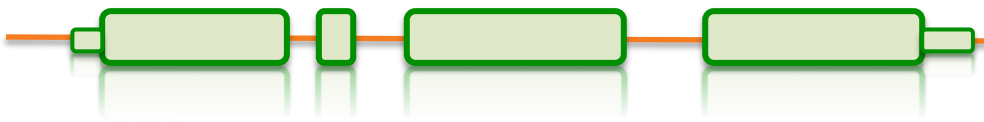
Interproscan

Sequence database	Version	Count	Count of proteins matching	
			any signature	integrated signatures
UniProtKB	2016_04	64237017	53062429 (82.6%)	51384050 (80.0%)
UniProtKB/TrEMBL	2016_04	63686057	52526761 (82.5%)	50852748 (79.8%)
UniProtKB/Swiss-Prot	2016_04	550960	535668 (97.2%)	531302 (96.4%)

InterPro2GO

Total number of GO terms mapped to InterPro entries - 31668

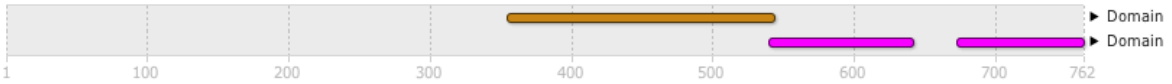
Interproscan results



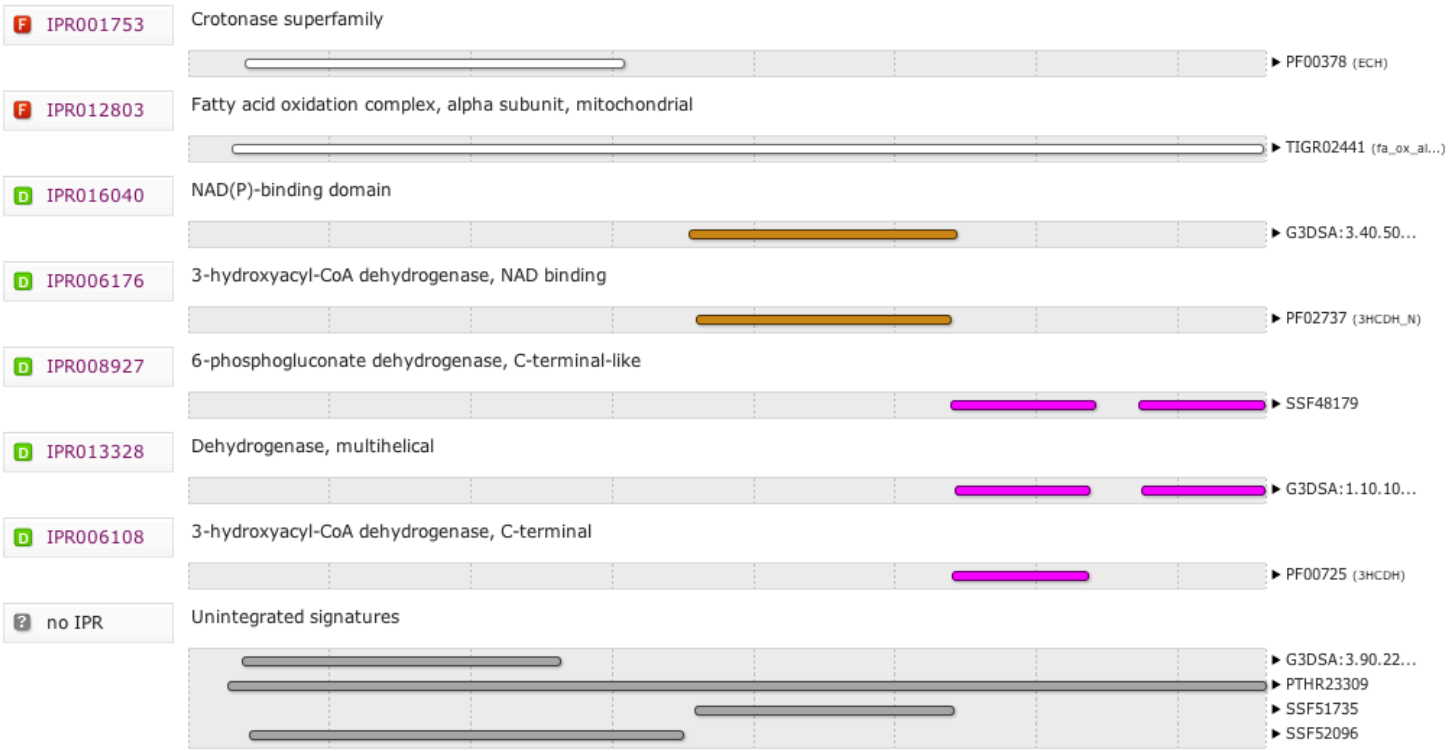
Protein family membership

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

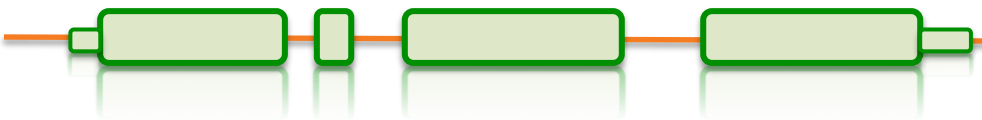
Domains and repeats



Detailed signature matches



Interproscan results



Ouput: TSV, XML, SVG, etc

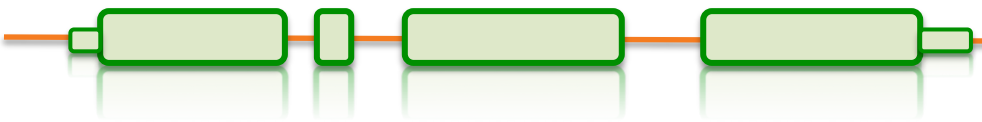
gene-2.44-mRNA-1 a9deba5837e2614a850c7849c85c8e9c 447 Pfam PF02458 Transferase family 98 425 1.4E-15 T 31-10-2015 IPR003480 Transferase GO:0016747

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

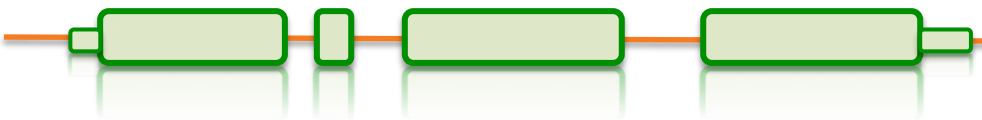
Interproscan-based approach



- Run Interproscan on the protein fasta file created by maker

- Use Maker-supplied scripts to merge the interproscan-results to the Maker annotations.gff file

Interproscan-based approach



- Looks for conserved domains, so might be more reliable than blast?

- How to go from conserved domains to assigning a function for your protein?

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

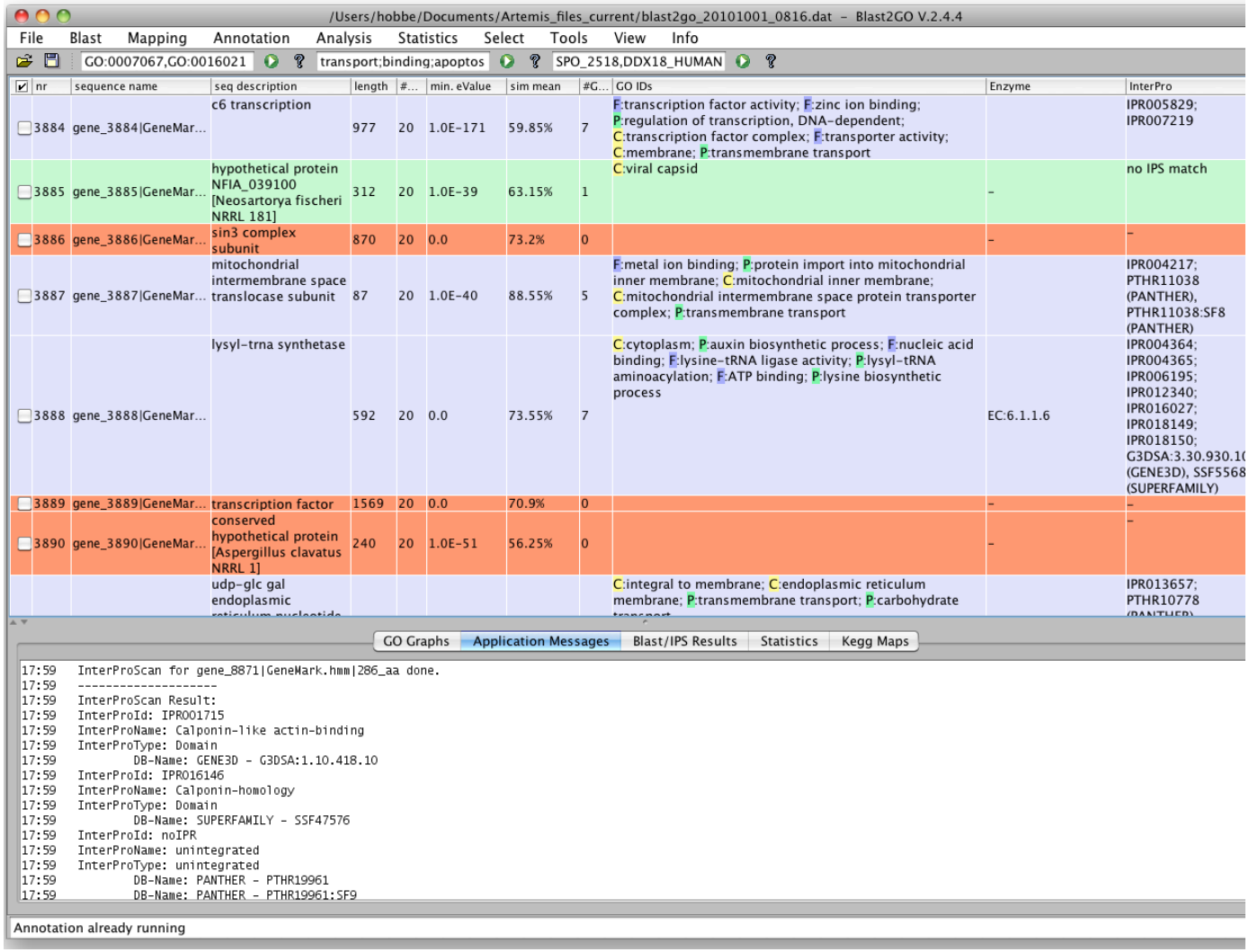
=> Contain nice downstream analysis/visualization components

Blast2GO

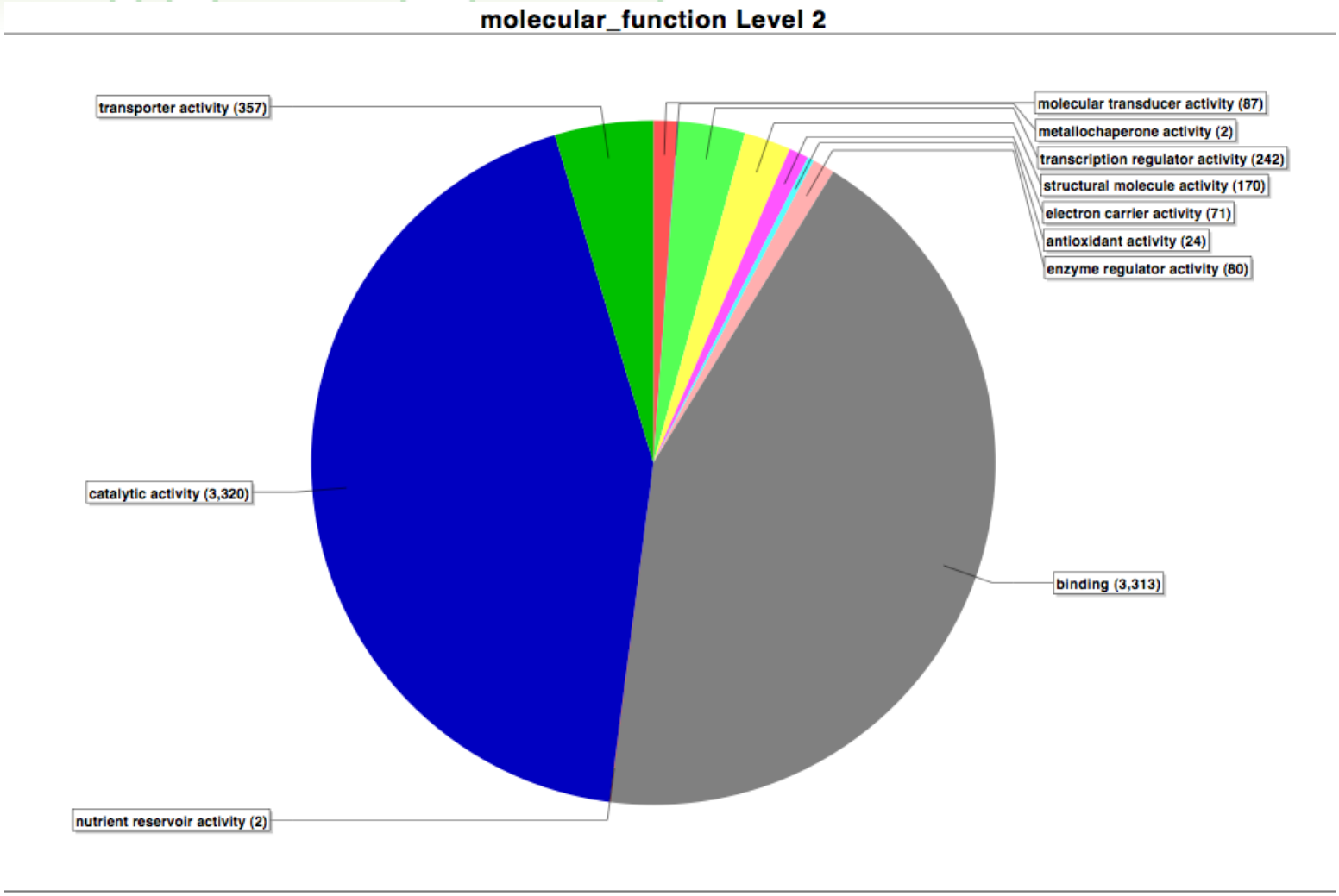
The NBIS annotation service

NATIONAL BIOINFORMATICS INFRASTRUCTURE SWEDEN

blast2go

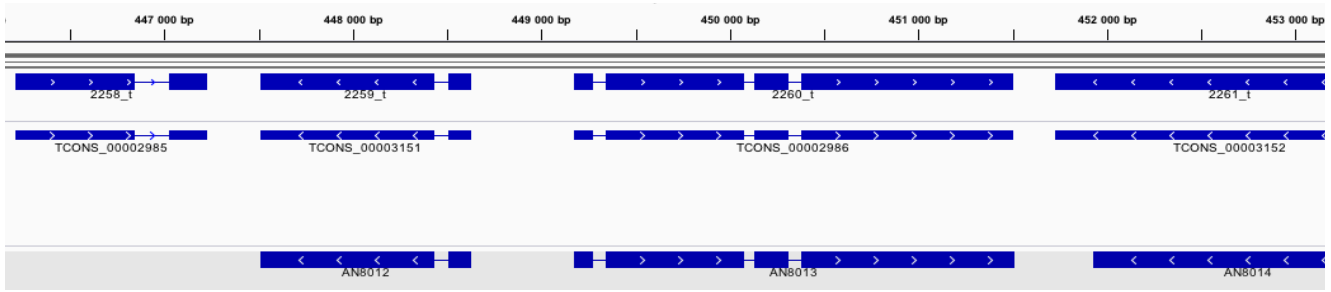


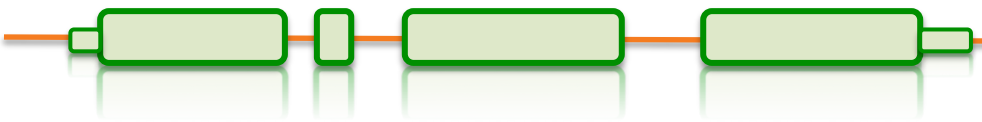
Blast2GO


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)

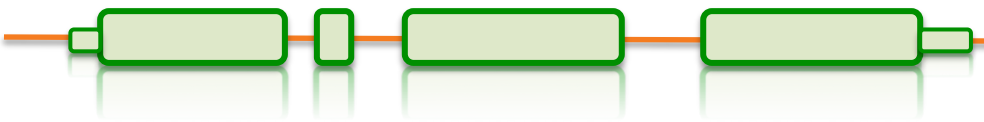
Liftovers are very useful for orthology determination

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





The END



supplement

KEGG-mapping