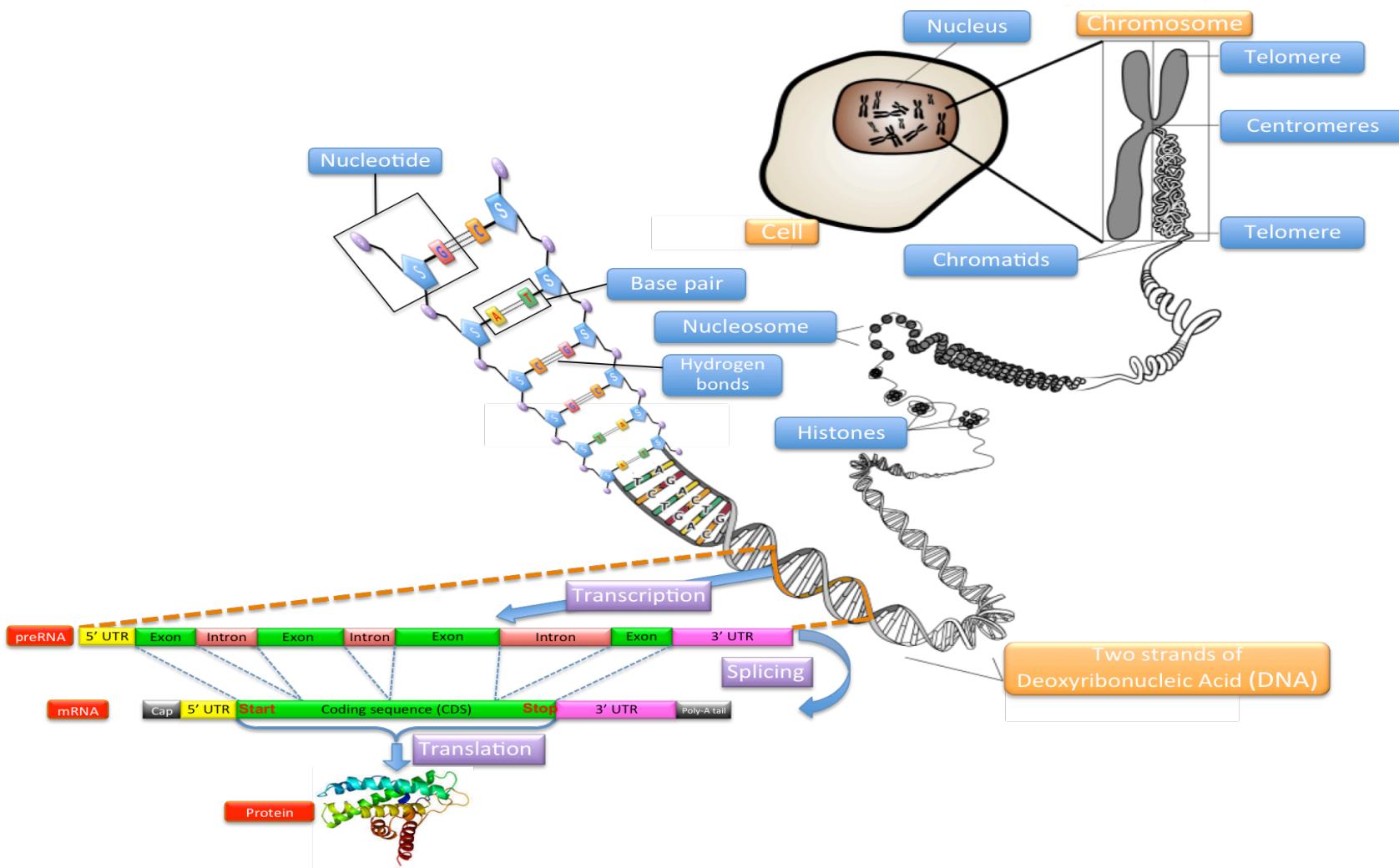
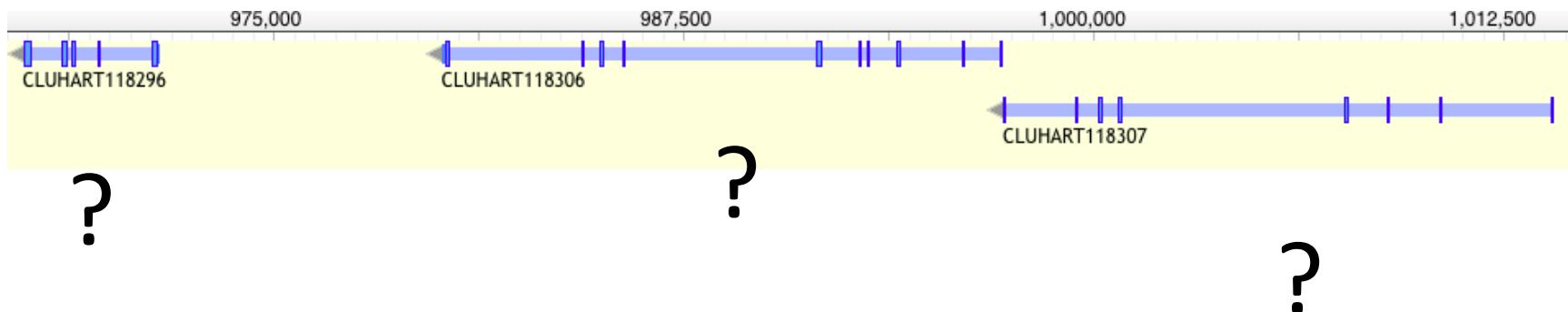


Functional annotation



The NBIS annotation service

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



Insulin receptor?

Vesicle-trafficking protein?

Alcohol dehydrogenase?

Aquaporin?

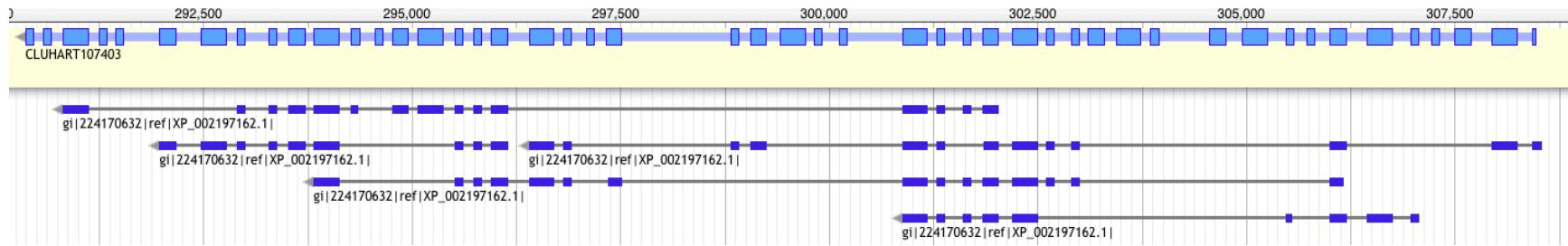
Transcription factor

MAP kinase kinase kinase?



The NBIS annotation service

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)

The NBIS annotation service

Methods

- Based on blast
 - Best blast hit
 - Clustering
- Based on synteny
 - ⇒ Whole genome alignment (lastZ)
 - (NBIS) Satsuma + kraken + custom script
- Based on phylogeny

Tools

The NBIS annotation service

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 !

Databases

The NBIS annotation service

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

More than 60 000 terms

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

Gene Ontology

The NBIS annotation service

Starting point of downstream analysis

The screenshot shows the homepage of the Gene Ontology Consortium (geneontology.org). The top navigation bar includes links for Home, Documentation, Downloads, Community, Tools, About, and Contact us. The main content area features a large search bar for "Search GO data". To the left, there's an "Enrichment analysis" tool with fields for gene IDs, ontology (set to biological process), species (Homo sapiens), and a "Submit" button. Below it are sections for "Statistics" (with a bar chart) and "Other GOC tools" (which is circled in orange). The right side contains sections for "Gene Ontology Consortium" (mission statement), "Search documentation" (with a search bar), "User stories" (with a link to personal user stories), and "What is the Gene Ontology?" (with a bulleted list of links). At the bottom, there's a "Recent news" section with a tweet about GO and a paper on extracellular RNA.

Gene Ontology Consortium

Home Documentation Downloads Community Tools About Contact us

Enrichment analysis

Your gene IDs here...

biological process

Homo sapiens

Submit

Advanced options / Help
Powered by PANTHER

Statistics

Other GOC tools

Explore other GOC tools in the AmiGO software suite.

Tweets about "#geneontology OR @news4go"

Recent news

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

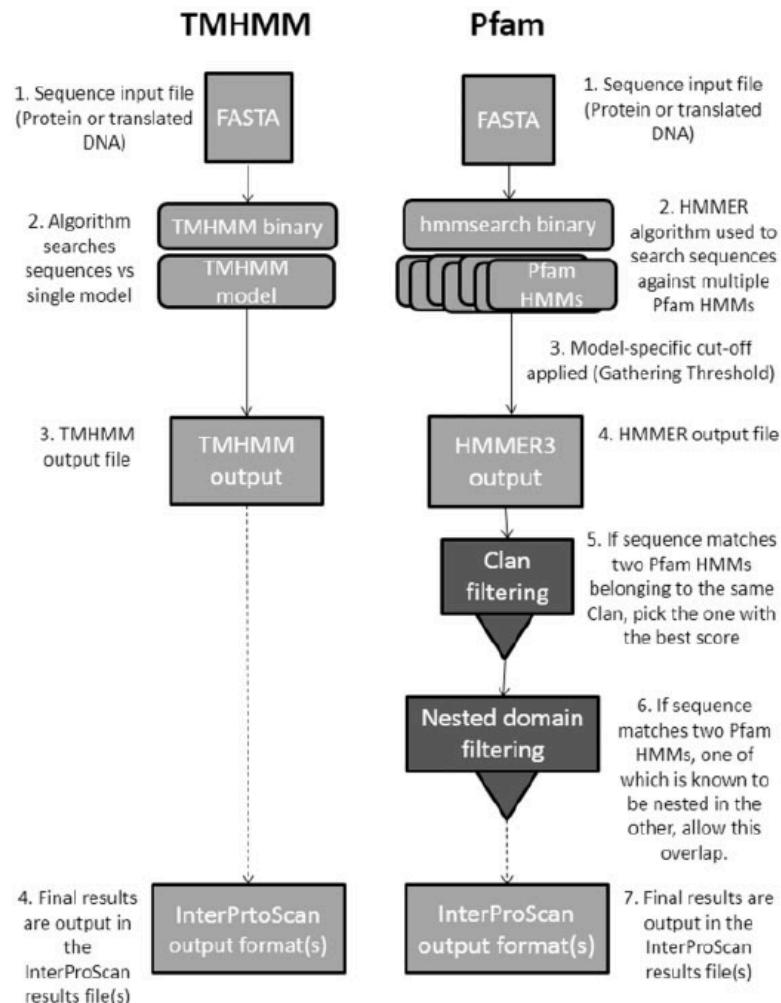
The NBIS annotation service

There is an equivalence table between GO and Interpro

GO annotation is given with an Evidence Code:

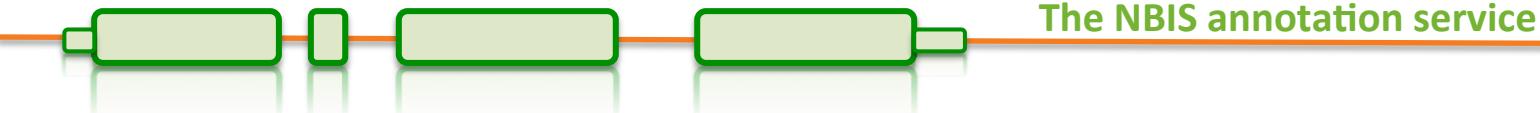
- IDA: inferred by direct assay
- ISS: inferred by sequence similarity
- IEA: electronic annotation
-

Search applications have two main modalities



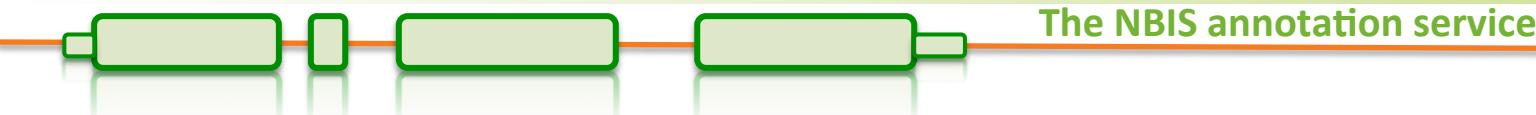
Jones,P.etal.InterProScan5:genome-scale protein function classification.
Bioinformatics 30, 1236–1240 (2014).

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



The NBIS annotation service

Strengths

- Fairly fast and easy
- Allow gene naming

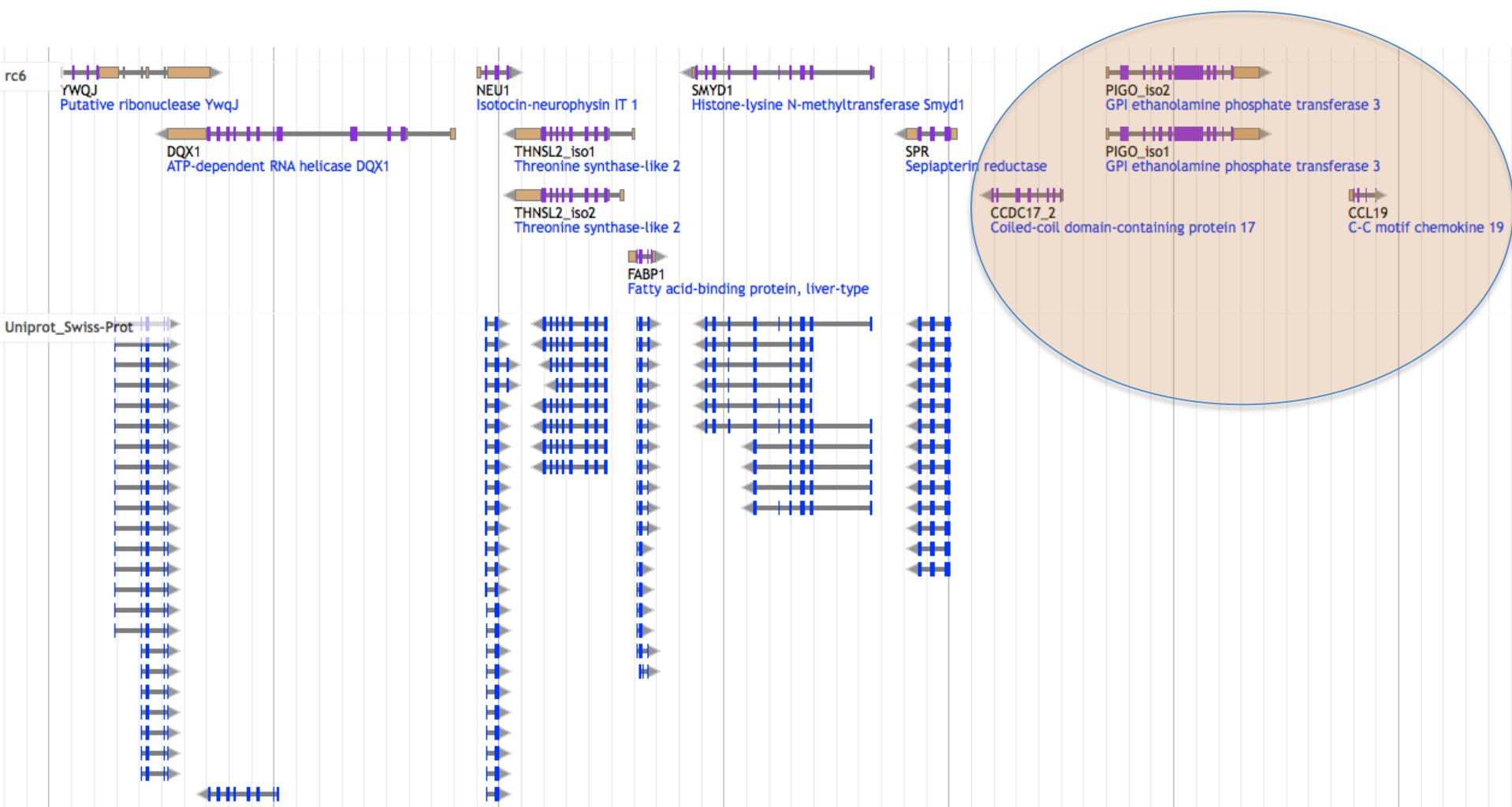
Limits

- 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

The NBIS annotation service

Aligned proteins ≠ Blast hits



Blast-based approach

The NBIS annotation service

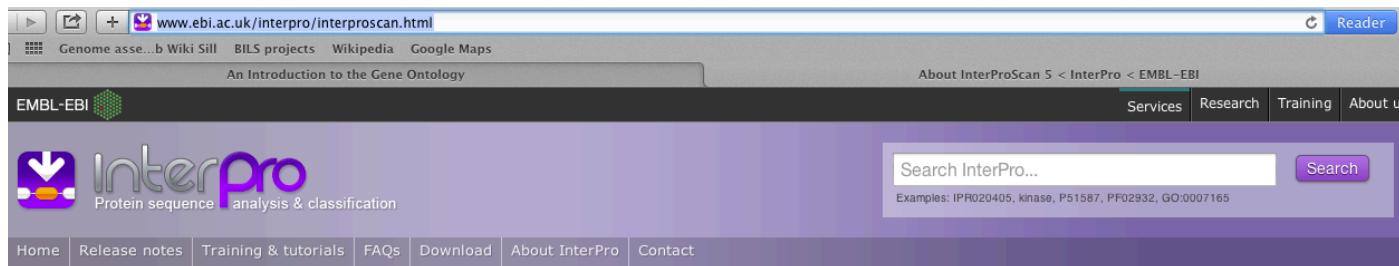


Blast-based annotation are tightly dependent to the quality of the annotation

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

The NBIS annotation service

- 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. Underneath the search bar, there's a note with examples: IPR020405, kinase, P51587, PF02932, GO:0007165. The main content area features the 'InterPro' logo and the tagline '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



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



Services

News
Brochures
Contact us
Intranet

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 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.

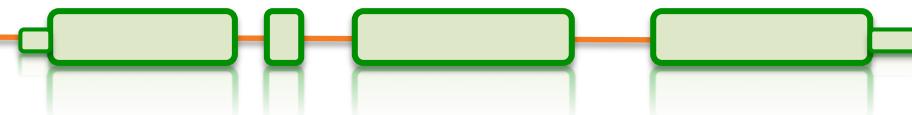
Member database information

	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

The NBIS annotation service



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%)

			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

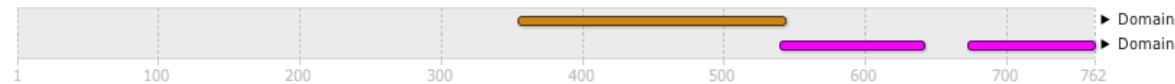
Interproscan results

Protein family membership

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

The NBIS annotation service

Domains and repeats



Detailed signature matches

■ IPR001753	Crotonase superfamily		► PF00378 (ECH)
■ IPR012803	Fatty acid oxidation complex, alpha subunit, mitochondrial		► TIGR02441 (fa_ox_al...)
■ IPR016040	NAD(P)-binding domain		► G3DSA:3.40.50...
■ IPR006176	3-hydroxyacyl-CoA dehydrogenase, NAD binding		► PF02737 (3HCDH_N)
■ IPR008927	6-phosphogluconate dehydrogenase, C-terminal-like		► SSF48179
■ IPR013328	Dehydrogenase, multihelical		► G3DSA:1.10.10...
■ IPR006108	3-hydroxyacyl-CoA dehydrogenase, C-terminal		► PF00725 (3HCDH)
■ no IPR	Unintegrated signatures		► G3DSA:3.90.22... ► PTHR23309 ► SSF51735 ► SSF52096

Interproscan results

Ouput: TSV, XML, SVG, etc

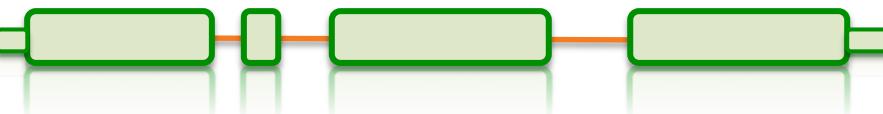
gene-2.44-mRNA-1 a9deba5837e2614a850c7849c85c8e9c 447 Pfam PF02458 Transferase family 98 425
1.4E-15 T 31-10-2015 IPR003480 Transferase GO:0016747

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

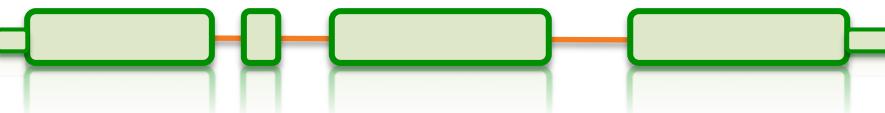
gene-1.4-mRNA-1 b867bbb377084bba6ea84dcda9f27f4e 511 Pfam PF07690 Major Facilitator Superfamily 67
447 3.5E-30 T 31-10-2015 IPR011701 Major facilitator superfamily GO:0016021|GO:0055085

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



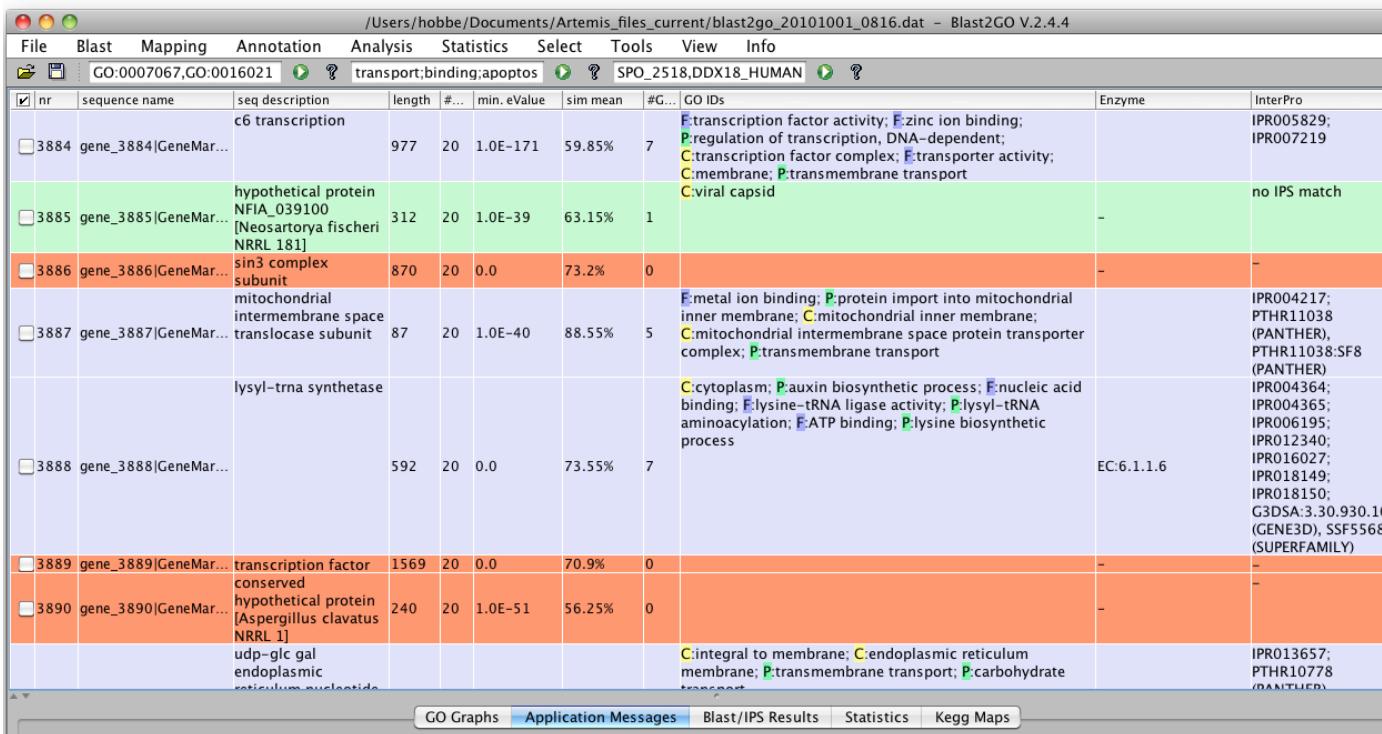
- Looks for conserved domains, so might be more reliable than blast?
- How to go from conserved domains to assigning a function for your protein?

The NBIS annotation service

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



A screenshot of the Blast2GO V.2.4.4 interface. The main window shows a search results table for gene_3884. The table includes columns for nr, sequence name, seq description, length, #..., min. eValue, sim mean, #G..., GO IDs, Enzyme, and InterPro. The GO IDs column contains a large amount of text describing biological processes. The Enzyme and InterPro columns also contain lists of identifiers. At the bottom of the table, there are tabs for GO Graphs, Application Messages, Blast/IPS Results, Statistics, and Kegg Maps.

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		IPR005829; IPR007219
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		IPR004217; 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	IPR004364; IPR004365; IPR006195; IPR012340; IPR016027; IPR018149; IPR018150; 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		IPR013657; 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: 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

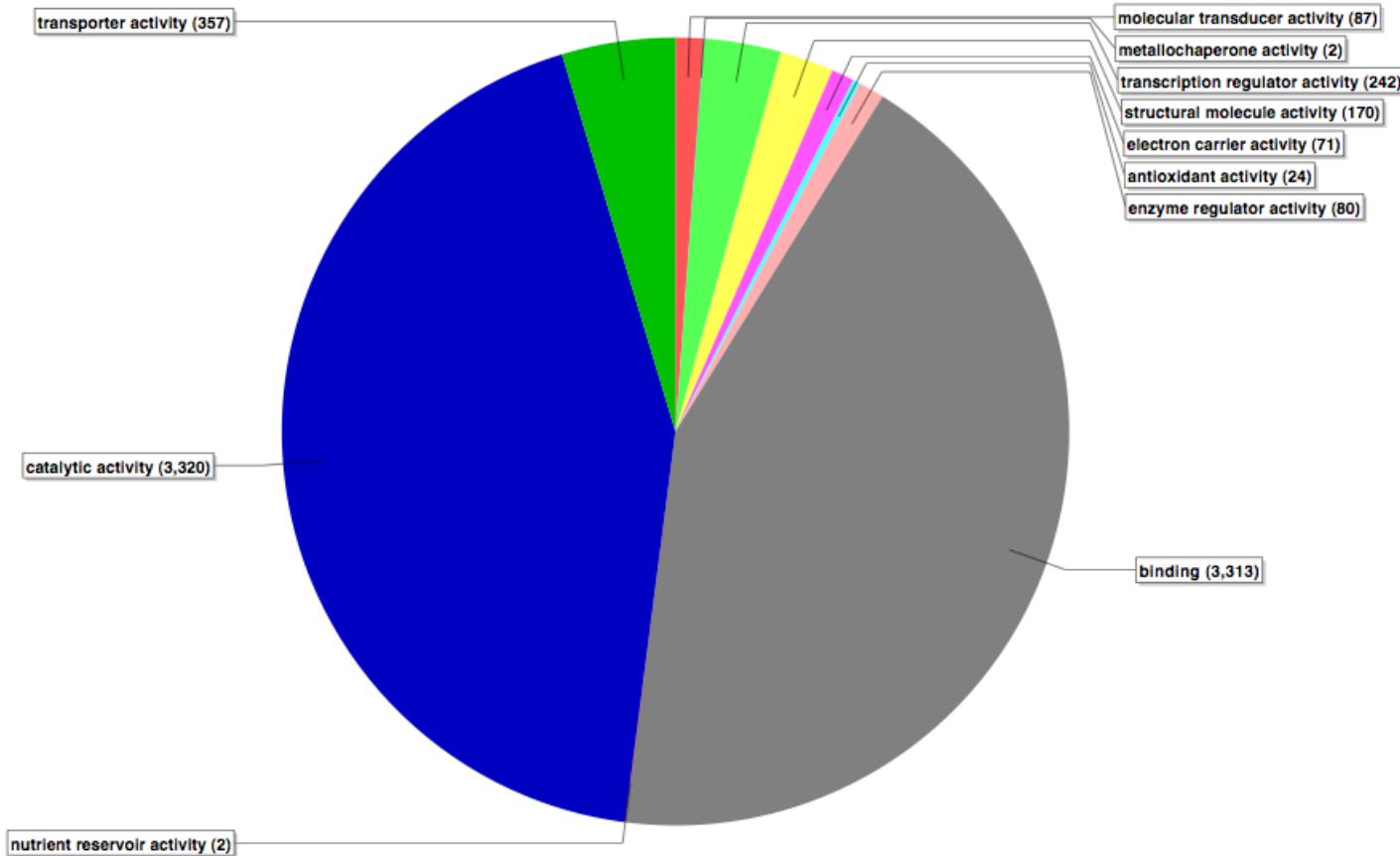
```

Annotation already running



blast2go

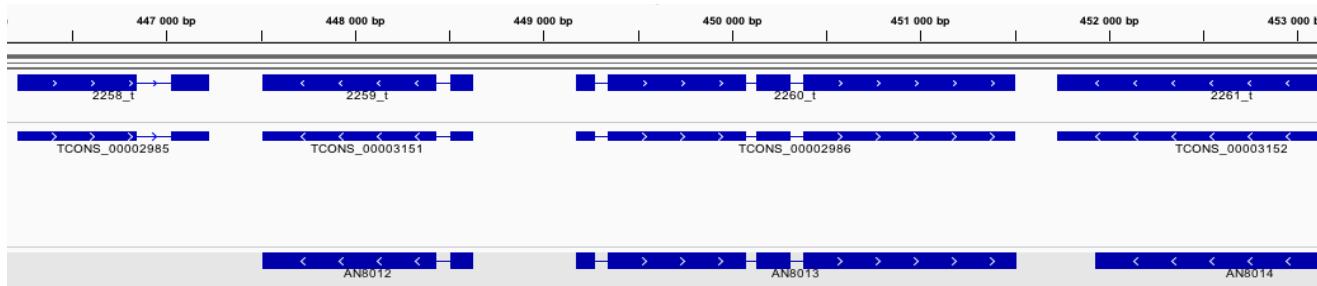
molecular_function Level 2

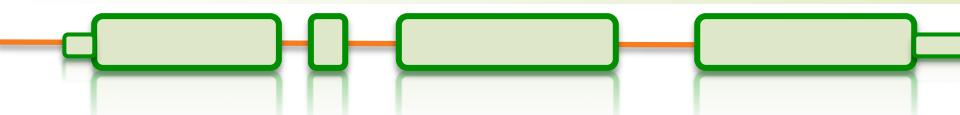


The NBIS annotation service

Liftovers are very useful for orthology determination

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





The NBIS annotation service

A diagram illustrating protein domain architecture. It consists of several rectangular boxes of varying sizes, each outlined in green, connected by a horizontal orange line. The boxes represent different domains, and their arrangement suggests a linear sequence or a domain fusion.

The END



The NBIS annotation service

supplement

KEGG-mapping

The NBIS annotation service

File Blast Mapping Annotation Analysis Statistics Select Tools View Help

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

succinyl- synthetase subunit ATP binding; F:succinate-CoA ligase (GDP-forming) activity; P:tricarboxylic acid cycle; C:succinate-CoA ligase IPR003781; IPR005810

GO Graphs Application Messages Blast/IPS Results Statistics Kegg Maps

GLYCEROLIPID METABOLISM

The diagram illustrates the biosynthesis of various glycerolipids from glycerol and fatty acids. Key intermediates include sn-Glycerol 3-phosphate, 1-Acyl-sn-glycerol 3-phosphate, 1,2-Diacyl-sn-glycerol 3-phosphate, and 1,2-Diacyl-3-(sn-glycero-1,2-diacylglycerol)sn-glycerol. Enzymes involved are color-coded: red (e.g., 1.1.1.2, 1.1.1.6), yellow (e.g., 2.3.1.15, 2.3.1.22), orange (e.g., 2.3.1.51, 2.3.1.23), green (e.g., 2.4.1.1, 2.4.1.2), blue (e.g., 2.7.1.30, 2.7.1.31), pink (e.g., 1.1.1.72), purple (e.g., 1.2.1.3), and light-red (e.g., 2.7.1.107).

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 Legend:

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