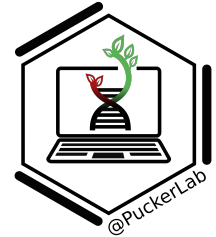
UNIVERSITÄT BONN



Prof. Dr. Boas Pucker

Pathway Databases

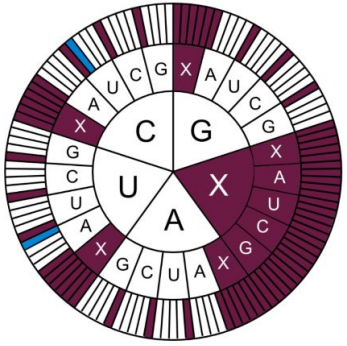
Availability of slides

- All materials are freely available (CC BY) - after the lectures:
  - eCampus: WBIO-A-08
  - GitHub: https://github.com/bpucker/teaching

- Questions: Feel free to ask at any time

- Feedback, comments, or questions: pucker[a]uni-bonn.de

Why do we need pathway databases?



Expanding the genetic code with an additional DNA base (X)

xanthosine 5'-phosphate (XMP) → guanosine 5'-phosphate (GMP)

guanosine monophosphate synthase (GMPS); EC6.3.5.2

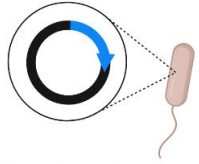
https://www.genome.jp/pathway/ec00230+6.3.5.2

Identification of candidate enzymes in a **database**

Croton tiglium is naturally producing a precursor of this unnatural DNA base

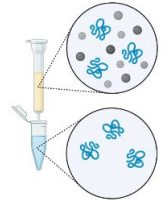
RNA extraction, sequencing, and construction of a transcriptome assembly

Cloning of GMPS coding sequence into a plasmid



Transformation into *Escherichia coli*

Heterologous expression of *GMPS*



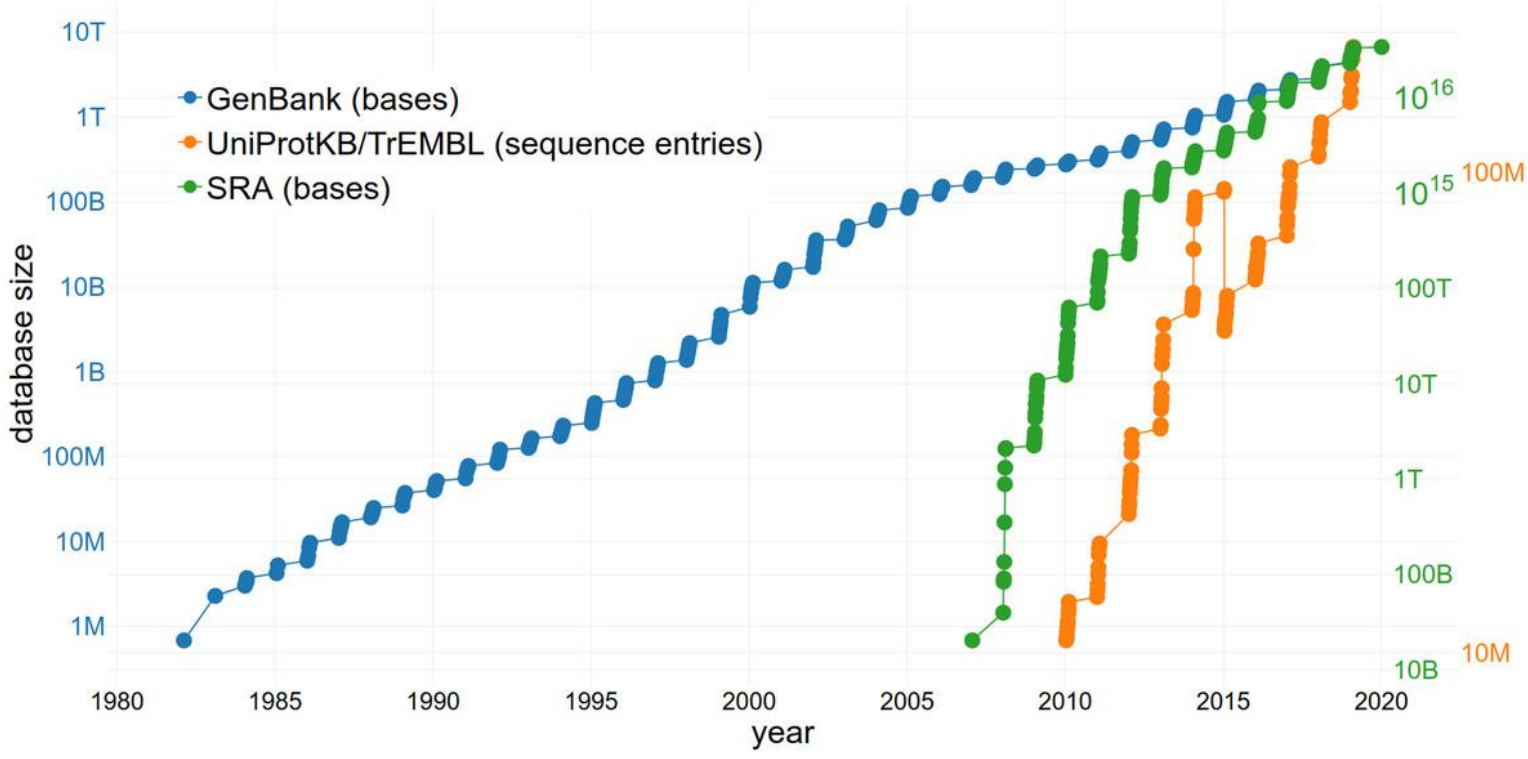
Protein purification

Enzyme assay

http://2017.igem.org/Team:Bielefeld-CeBiTec
Haak et al., 2018: 10.3389/fmolb.2018.00062

UNIVERSITÄT BONN

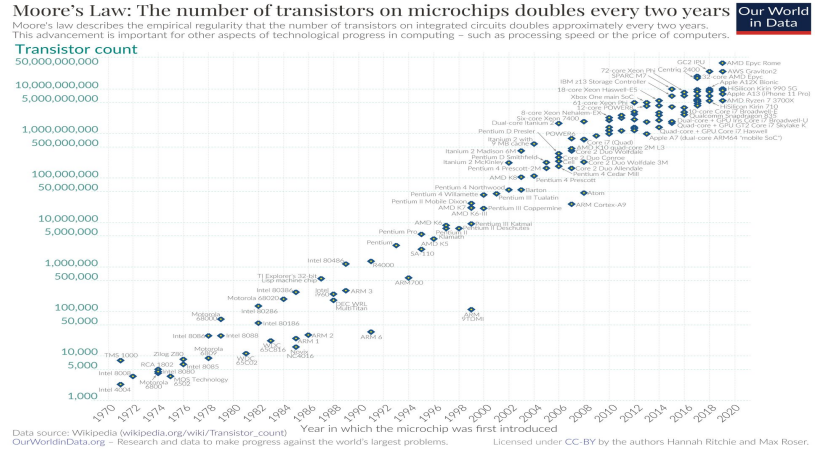
Exponential growth of databases



Sielemann & Hafner et al., 2020: 10.7717/peerj.9954

Generation of large data sets

- Sequencing technologies are outpacing Moore's law
  - Moore's law = number of transistors on microchips doubles every two years



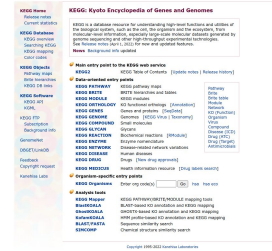
https://commons.wikimedia.org/wiki/File:Moore%27s_Law_Transistor_Count_1970-2020.png

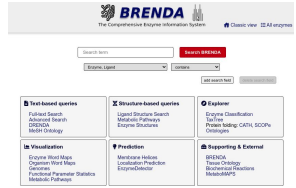
- Sequencing is becoming widely available (democratization of sequencing)

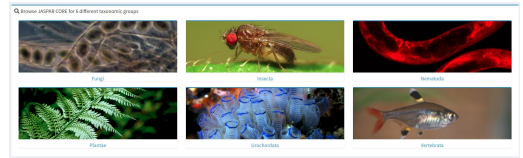
- Proteomics and metabolomics are developing rapidly

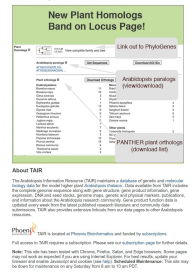
Types of pathway databases

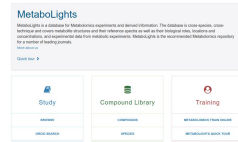
- Reactions (KEGG, BioCyc, MetaCyc, Reactome)
- Enzymes (BRENDA)
- Transcription factors (JASPAR)
- Metabolites (MetaboLights)
- Species specific databases (TAIR, Phytozome, PLAZA)











https://www.genome.jp/kegg/; https://www.brenda-enzymes.org/
https://www.ebi.ac.uk/metabolights/; https://www.arabidopsis.org/

UNIVERSITÄT BONN

How to access data?

- Open Data = Initiative to make research data publicly available

https://www.go-fair.org/

- FAIR data = Findable Accessible Interoperable Reuseable data
  - Findable = registered in a central database
  - Accessible = available through a repository (not just from authors)
  - Interoperable = data format allows automatic processing
  - Reusable = sufficient metadata to use data sets in other studies

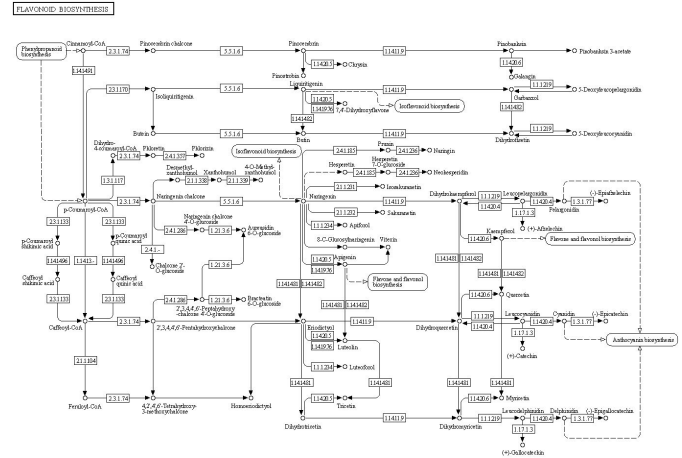
- ORCID is used as a universal login solution across many platforms
  - ORCID = Open Researcher and Contributor ID
  - SSO = Single Sign-On (user stays logged in across websites)

Connecting research and researchers

https://orcid.org/

KEGG: Kyoto Encyclopedia of Genes and Genomes

- Maps of pathways showing the individual reactions with catalyzing enzymes
- Information about genomes and genes
- Chemical details about enzymes, substrates, and products
- KEGG is financed through a subscription model (for FTP download), but website is freely accessible

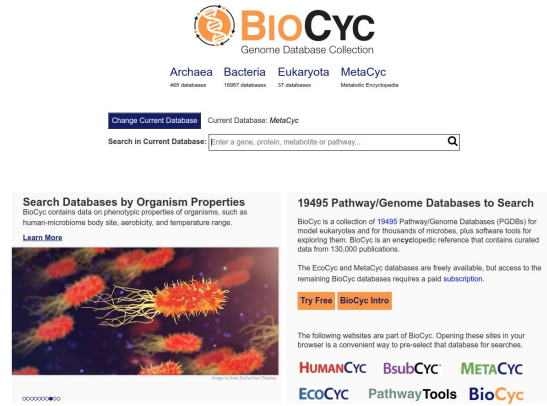


map00941 Flavonoid biosynthesis

Kanehisa & Goto, 2000: https://doi.org/10.1093/nar/28.1.27
Kanehisa et al., 2022: https://doi.org/10.1002/pro.4172; https://www.genome.jp/kegg/

BioCyc: Collection of databases

- Integration of databases and tools for omics
- Curation is involved to increase the quality
- Strong focus on microorganism:
  - EcoCyc = **E**scherichia **co**li En**cyc**lopedia
  - BsubCyc = **B**acillus **sub**tilis En**cyc**lopedia
  - YeastCyc = **Yeast** En**cyc**lopedia
- BUT:
  - Search for metabolites
  - Search for pathways
  - Search for reactions

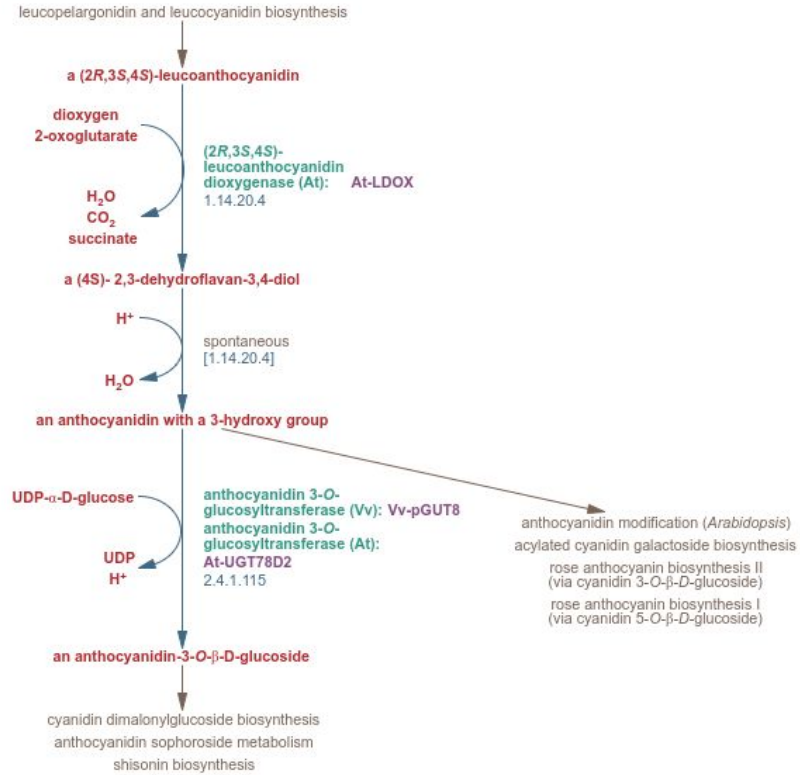


Karp et al., 2019: 10.1093/bib/bbx085
https://biocyc.org/

UNIVERSITÄT BONN

MetaCyc: Metabolite Encyclopedia

- Integrates genomic data with functional annotation
- Visualization of pathway databases
- Shows intermediates and enzymes of biosynthesis pathways
- MetaFlux: flux-balance analysis



https://metacyc.org/; Karp et al., 2015: https://doi.org/10.48550/arXiv.1510.03964
Figure: https://biocyc.org/META/NEW-IMAGE?type=PATHWAY&object=PWY-5125

UNIVERSITÄT BONN

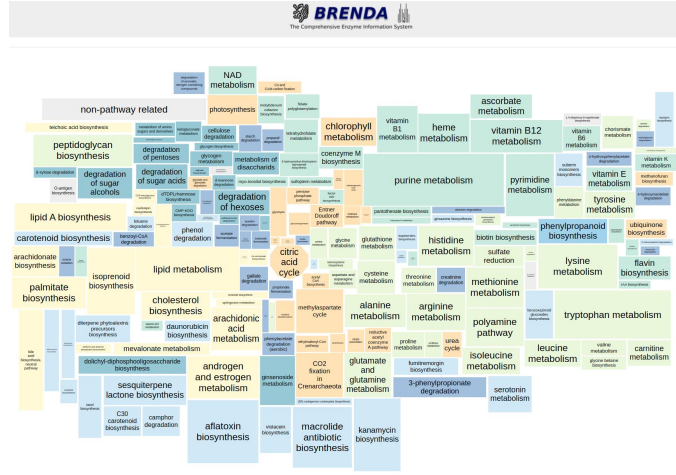
Enzyme Commission (EC) numbers

- Four digit system to classify enzymes:
  - General type of reaction
  - Subclass
  - Sub-subclass
  - Serial number (within sub-subclass)
- General types:
  - EC1 = Oxidoreductases
  - EC2 = Transferases
  - EC3 = Hydrolases
  - EC4 = Lyases
  - EC5 = Isomerases
  - EC6 = Ligases

https://enzyme.expasy.org/

BRENDA: BRaunschweig ENzyme DAtabase

- Enzyme database hosted at TU Braunschweig (BRICS)
- Text and structure-based queries
- Visualization of pathways
- Manual curation of datasets
- Many details about enzyme properties (substrates, kinetics, mutants, …)



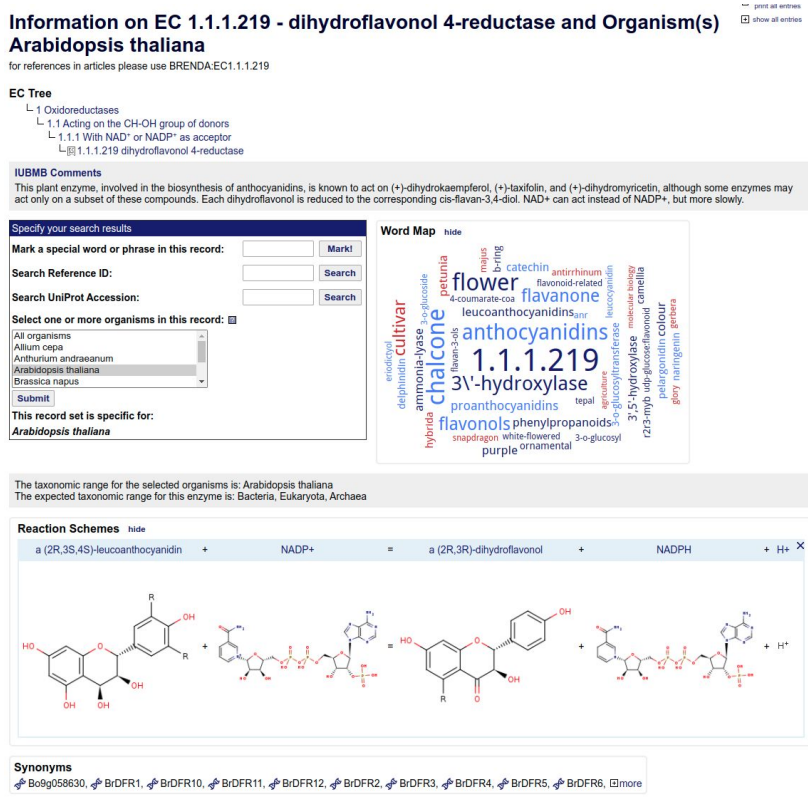
www.brenda-enzymes.org
Chang et al., 2021: 10.1093/nar/gkaa1025

BRENDA - example (1)

- DFR (dihydroflavonol 4-reductase) is a central enzyme in the anthocyanin biosynthesis

- BRENDA provides information about reaction mechanism and species-specific parameters

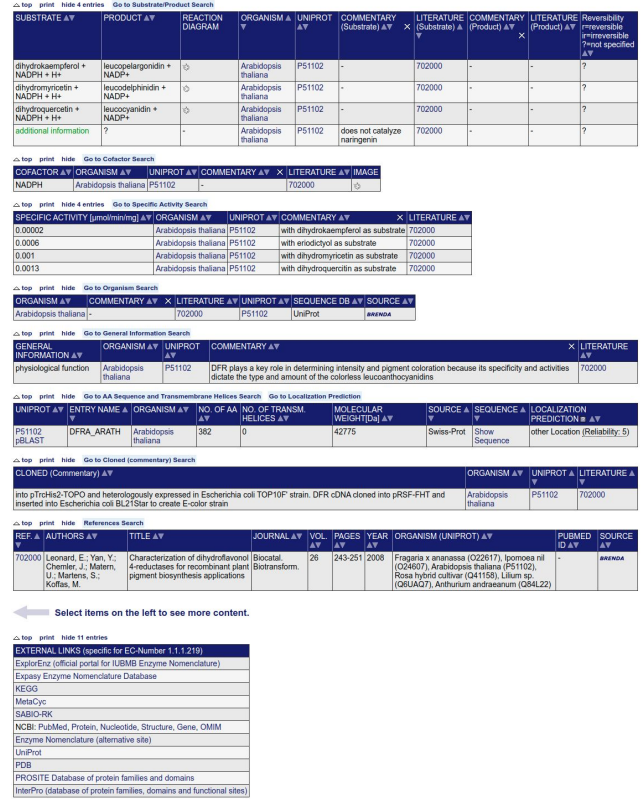
- Enzymatic properties of multiple species allow quick comparison



https://www.brenda-enzymes.org/

BRENDA - example (2)

- Specific substrates and products are listed
- K_M and V_{max} values are included
- References point to publications and other databases



top print hide 4 entries Go to Substrate/Product Search

SUBSTRATE	PRODUCT	REACTION DIAGRAM	ORGANISM	UNIPROT	COMMENTARY (Substrate)	LITERATURE (Substrate)	COMMENTARY (Product)	LITERATURE (Product)	Reversibility r=reversible ir=irreversible ?=not specified
dihydrokaempferol + NADPH + H+	leucopelargonidin + NADP+		Arabidopsis thaliana	P51102	-	702000	-	-	?
dihydromyricetin + NADPH + H+	leucodelphinidin + NADP+		Arabidopsis thaliana	P51102	-	702000	-	-	?
dihydroquercetin + NADPH + H+	leucocyanidin + NADP+		Arabidopsis thaliana	P51102	-	702000	-	-	?
additional information	?	-	Arabidopsis thaliana	P51102	does not catalyze naringenin	702000	-	-	?

top print hide Go to Cofactor Search

COFACTOR	ORGANISM	UNIPROT	COMMENTARY	LITERATURE	IMAGE
NADPH	Arabidopsis thaliana	P51102	-	702000	

top print hide 4 entries Go to Specific Activity Search

SPECIFIC ACTIVITY [μmol/min/mg]	ORGANISM	UNIPROT	COMMENTARY	LITERATURE
0.00002	Arabidopsis thaliana	P51102	with dihydrokaempferol as substrate	702000
0.0006	Arabidopsis thaliana	P51102	with eriodictyol as substrate	702000
0.001	Arabidopsis thaliana	P51102	with dihydromyricetin as substrate	702000
0.0013	Arabidopsis thaliana	P51102	with dihydroquercitin as substrate	702000

top print hide Go to Organism Search

ORGANISM	COMMENTARY	LITERATURE	UNIPROT	SEQUENCE DB	SOURCE
Arabidopsis thaliana	-	702000	P51102	UniProt	BRENDA

top print hide Go to General Information Search

GENERAL INFORMATION	ORGANISM	UNIPROT	COMMENTARY	LITERATURE
physiological function	Arabidopsis thaliana	P51102	DFR plays a key role in determining intensity and pigment coloration because its specificity and activities dictate the type and amount of the colorless leucoanthocyanidins	702000

top print hide Go to AA Sequence and Transmembrane Helices Search Go to Localization Prediction

UNIPROT	ENTRY NAME	ORGANISM	NO. OF AA	NO. OF TRANSM. HELICES	MOLECULAR WEIGHT [Da]	SOURCE	SEQUENCE	LOCALIZATION PREDICTION
P51102 pBLAST	DFRA_ARATH	Arabidopsis thaliana	382	0	42775	Swiss-Prot	Show Sequence	other Location (Reliability: 5)

top print hide Go to Cloned (commentary) Search

CLONED (Commentary)	ORGANISM	UNIPROT	LITERATURE
into pTrcHis2-TOPO and heterologously expressed in Escherichia coli TOP10F' strain. DFR cDNA cloned into pRSF-FHT and inserted into Escherichia coli BL21Star to create E-color strain	Arabidopsis thaliana	P51102	702000

top print hide References Search

REF.	AUTHORS	TITLE	JOURNAL	VOL.	PAGES	YEAR	ORGANISM (UNIPROT)	PUBMED ID	SOURCE
702000	Leonard, E.; Yan, Y.; Chemler, J.; Matern, U.; Martens, S.; Koffas, M.	Characterization of dihydroflavonol 4-reductases for recombinant plant pigment biosynthesis applications	Biocatal. Biotransform.	26	243-251	2008	Fragaria x ananassa (O22617), Ipomoea nil (O24607), Arabidopsis thaliana (P51102), Rosa hybrid cultivar (Q41158), Lilium sp. (Q6UAQ7), Anthurium andraeanum (Q84L22)	-	BRENDA

Select items on the left to see more content.

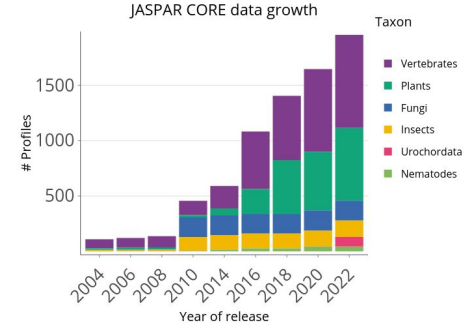
top print hide 11 entries

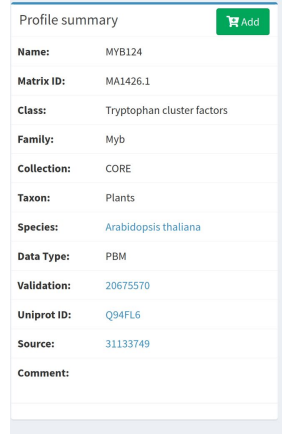
EXTERNAL LINKS (specific for EC-Number 1.1.1.219)
ExplorEnz (official portal for IUBMB Enzyme Nomenclature)
Expasy Enzyme Nomenclature Database
KEGG
MetaCyc
SABIO-RK
NCBI: PubMed, Protein, Nucleotide, Structure, Gene, OMIM
Enzyme Nomenclature (alternative site)
UniProt
PDB
PROSITE Database of protein families and domains
InterPro (database of protein families, domains and functional sites)

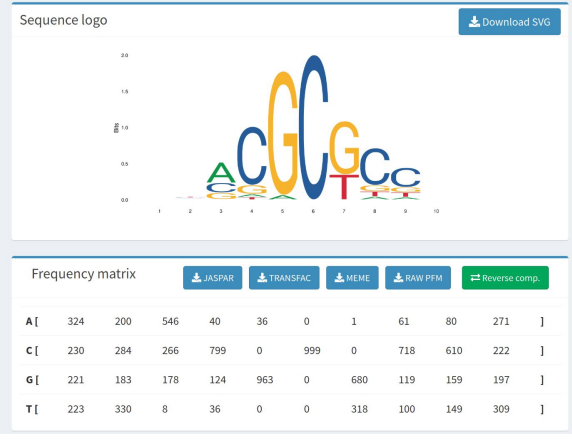
https://www.brenda-enzymes.org/

JASPAR: THE transcription factor database

- Contains motifs of transcription factor binding sites
- Additional details about transcription factors
- Most comprehensive database
- Regularly updated



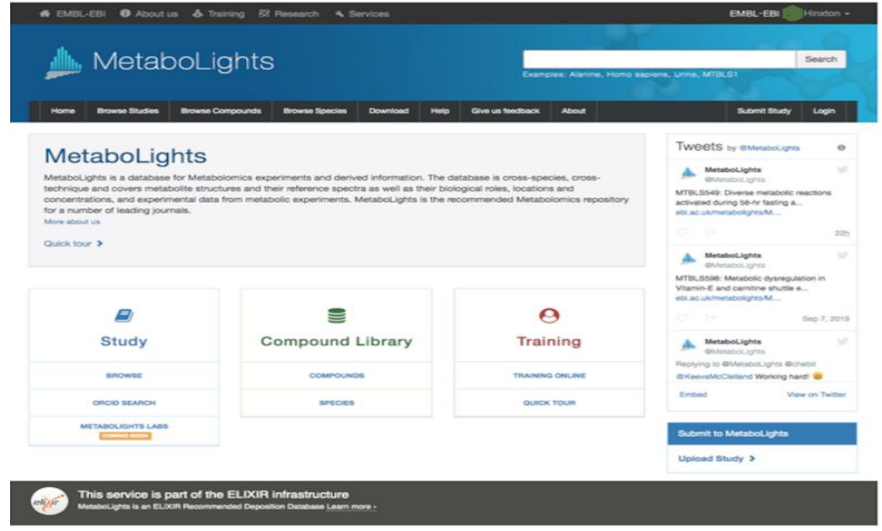




UNIVERSITÄT BONN

MetaboLights

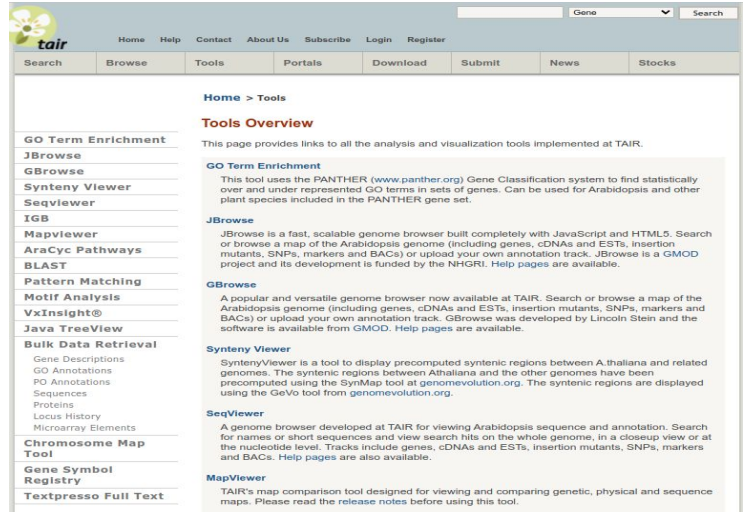
- Database of metabolomics experiments
- Hosted at European Bioinformatics Institute (UK)
- Collaboration with journals to enforce data submission



https://www.ebi.ac.uk/metabolights/
Haug et al., 2019: https://doi.org/10.1093/nar/gkz1019

Species specific databases: TAIR

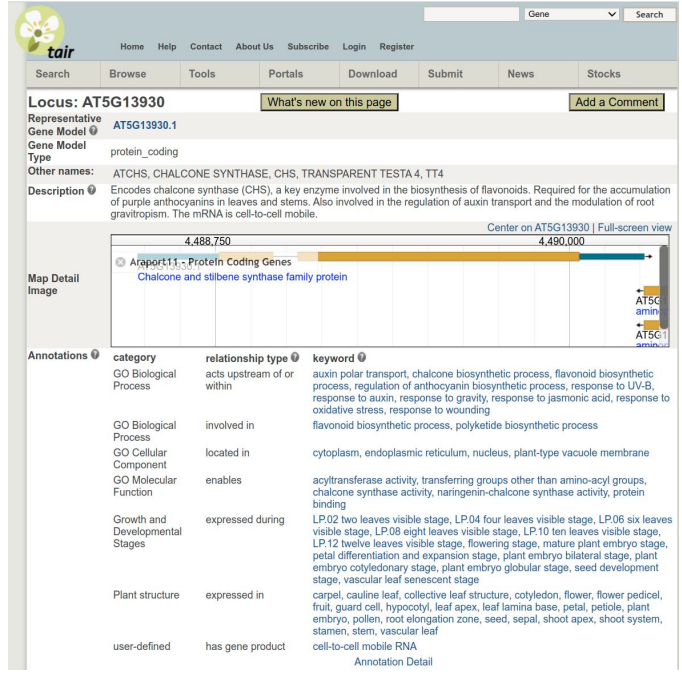
- TAIR is located at Phoenix Bioinformatics
- Institutional subscriptions provide funding for maintenance
- Full access requires subscription, but limited number of visits is free
- Collection of data sets and tools (e.g. BLAST)
- Current version: TAIR10 (=2010)
- Substitution by Araport11 was started, but is no longer funded



https://www.arabidopsis.org/index.jsp

TAIR10 - functional annotation

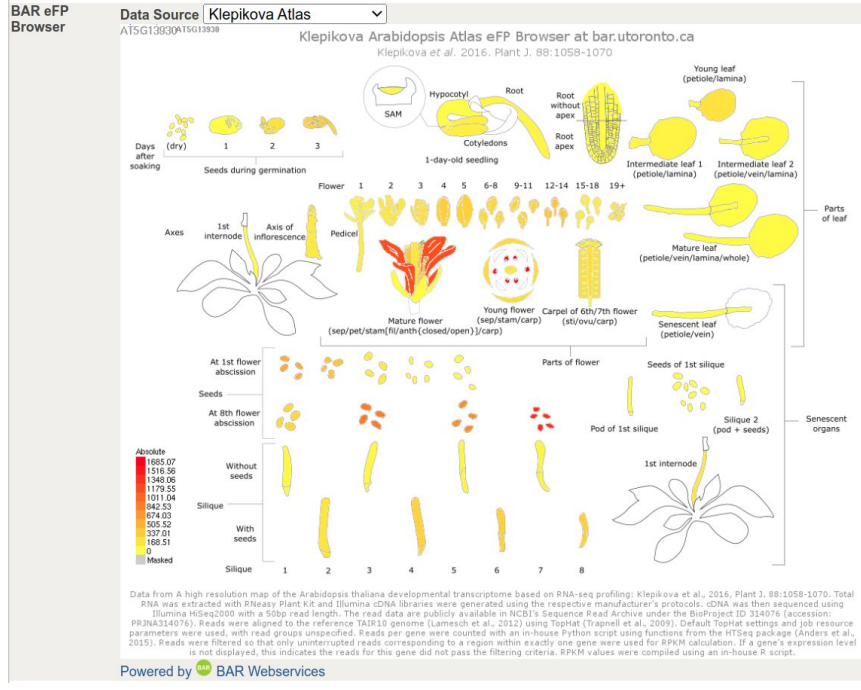
- TAIR10 provides information about function of genes
- References to other database for additional information
- Lists of publications about a given gene
- *Arabidopsis thaliana* gene functions are transferred to genes of other species



https://www.arabidopsis.org/

TAIR10 - eFP Browser

- Gene expression analysis is integrated in TAIR
- eFP browser provides expression data
- eFP = electronic Fluorescent Pictograph
- Microarray and RNA-Seq data sets are basis of gene expression analysis



http://bar.utoronto.ca/efp2/Arabidopsis/Arabidopsis_eFPBrowser2.html
Winter et al., 2007: 10.1371/journal.pone.0000718

UNIVERSITÄT BONN

TAIR10 - references

- Aggregates all references reporting about a gene
- Connection of knowledge
- AGI = Arabidopsis Gene Identifier
- AGIs mentioned in publication are used to assign studies to genes

Publication	author/title	source	associated loci	date
	Dahhan, D A, Reynolds, G D, Cár… Proteomic characterization of isolated Arabidopsis clathrin-coated vesicles reveals evolutionarily conserved and plant-specific components	THE PLANT CELL	AT2G46830 AT3G01780 AT3G51240 AT5G13930 AT5G61380	2022
	Verma, D, Bhagat, P K, Sinha, A K A dual-specificity phosphatase, MAP kinase phosphatase 1, positively regulates blue light-mediated seedling development in Arabidopsis	PLANTA	AT1G29930 AT1G67090 AT2G04550 AT2G43790 AT2G46340 AT3G06110 AT3G23610 AT3G55270 AT4G36730 AT5G11260 AT5G13930 AT5G23720 AT5G24120 AT5G40440	2021
	Kiselev, K V, Suprun, A R, Aleyno… External dsRNA Downregulates Anthocyanin Biosynthesis-Related Genes and Affects Anthocyanin Accumulation in Arabidopsis thaliana	INTERNATIONAL JOURNAL OF MOLECULAR SCIENCES	AT1G71030 AT1G77450 AT5G13930	2021
	Markus, C, Pecinka, A, Merotto, A Insights into the Role of Transcriptional Gene Silencing in Response to Herbicide-Treatments in Arabidopsis thaliana	INTERNATIONAL JOURNAL OF MOLECULAR SCIENCES	AT2G36490 AT5G13930	2021
	Zhang, X, He, Y, Li, L, Liu, H, Hon… Involvement of the R2R3-MYB transcription factors MYB21 and its homologs in regulating the stamen flavonols accumulation in Arabidopsis	JOURNAL OF EXPERIMENTAL BOTANY	AT1G30135 AT2G47460 AT3G01530 AT3G27810 AT3G62610 AT5G08640 AT5G13930 AT5G40348 AT5G40350 AT5G49330 FLS1	2021
	Yu, Z C, Zheng, X T, Lin, W, He,… Photoprotection of Arabidopsis leaves under short-term high light treatment: The antioxidant capacity is more important than the anthocyanin shielding effect	PLANT PHYSIOLOGY AND BIOCHEMISTRY	AT1G17610 AT4G16860 AT4G22880 AT5G13930	2021
	Chapman, J M, Muday, G K Flavonols modulate lateral root emergence by scavenging reactive oxygen species in Arabidopsis thaliana	JOURNAL OF BIOLOGICAL CHEMISTRY	AT5G07990 AT5G08640 AT5G13930	2020
	Gayomba, S R, Muday, G K Flavonols regulate root hair development by modulating accumulation of reactive oxygen species in the root epidermis	DEVELOPMENT	AT5G13930 AT5G51060	2020
	Weng, C Y, Zhu, M H, Liu, Z Q, Z… Integrated bioinformatics analyses identified SCL3-induced regulatory network in Arabidopsis thaliana roots	BIOTECHNOLOGY LETTERS	AT1G50420 AT3G51240 AT3G55120 AT5G05270 AT5G13930	2020
	Naake, T, Maeda, H A, Proost, S,… Kingdom-wide analysis of the evolution of the plant type III polyketide synthase superfamily	PLANT PHYSIOLOGY	AT1G02050 AT4G34850 AT5G13930	2020
	Hamamouch, N, Winkel, B S J, Li… Modulation of Arabidopsis Flavonol Biosynthesis Genes by Cyst and Root-Knot Nematodes	PLANTS (BASEL)	AT2G47460 AT5G08640 AT5G13930	2020
	Cheng, W, Lin, M, Qiu, M, Kong,… Chitin synthase is involved in vegetative growth, asexual reproduction, and pathogenesis of Phytophthora capsici and P. sojae	ENVIRONMENTAL MICROBIOLOGY	AT2G33580 AT3G21630 AT5G13930	2019
	Nakayama, T, Takahashi, S, Waki, T Formation of Flavonoid Metabolons: Functional Significance of Protein-Protein Interactions and Impact on Flavonoid Chemodiversity	FRONT PLANT SCI	AT5G13930	2019
	Rai, N, Neugart, S, Yan, Y, Wang,… How do cryptochromes and UVR8 interact in natural and simulated sunlight?	JOURNAL OF EXPERIMENTAL BOTANY	AT5G13930 AT5G63860	2019

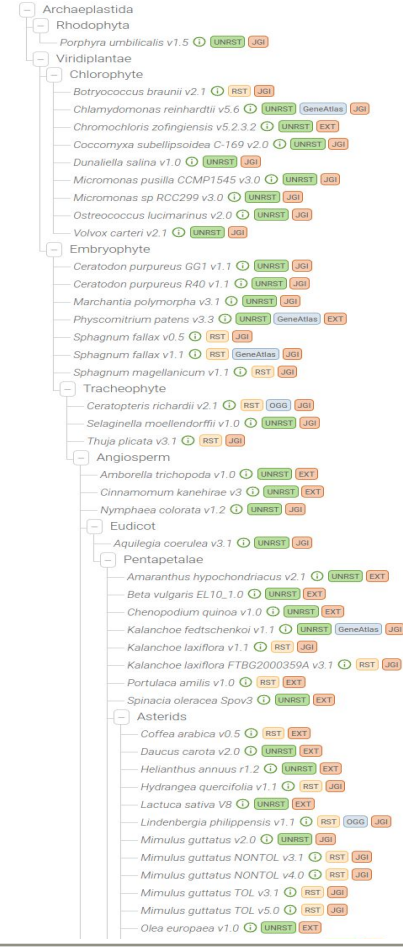
View Complete List (15 of 96 displayed)

https://www.arabidopsis.org/

UNIVERSITÄT BONN

Phytozome

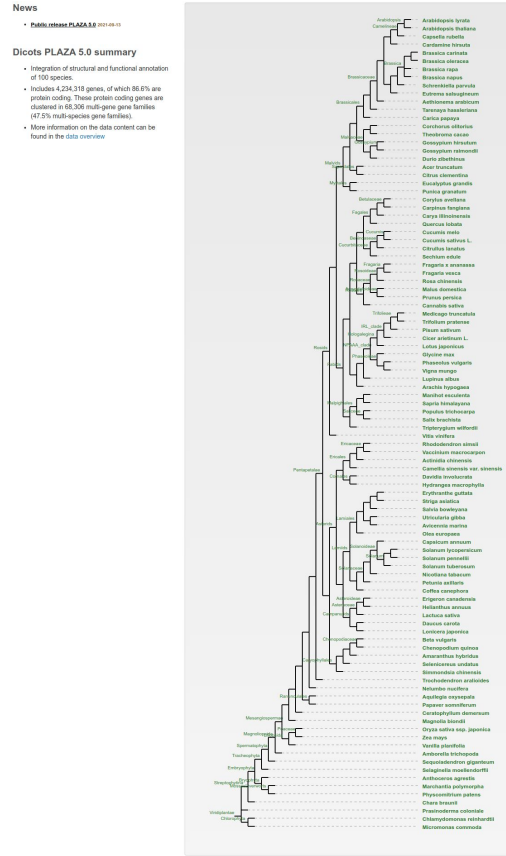
- Operated by the Joint Genome Institute (JGI), USA
- Collection of plant genome sequences and the corresponding annotations (best annotation source!)
- Data sets are available for download
- V13 contains 261 genome assemblies
- References of all data sets are clearly presented



https://phytozome-next.jgi.doe.gov/
Goodstein et al., 2012: 10.1093/nar/gkr944

UNIVERSITÄT BONN PLAZA

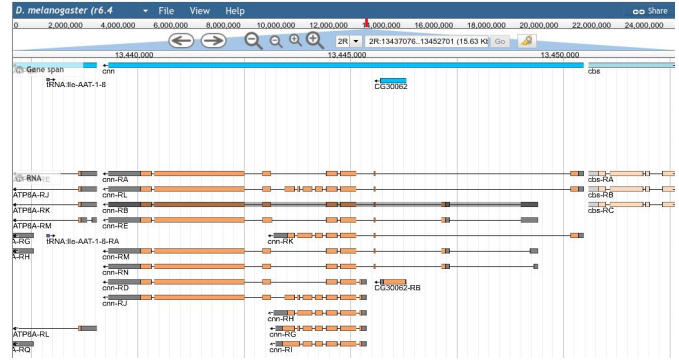
- Hosted in Belgium (Department of Plant Biotechnology and Bioinformatics); partly funded by ELIXIR
- Collection of plant genome sequences and corresponding annotations
- Focus on comparative genomics
- PLAZA v5 contains 134 genome sequences and their annotations



https://bioinformatics.psb.ugent.be/plaza/
Van Bel et al., 2021: 10.1093/nar/gkab1024

UNIVERSITÄT BONN FlyBase

- Website of the Drosophila research community
- BLAST service
- JBrowse = Web service to inspect the genome sequence
- Gene expression data; Data downloads
- Information about meetings, job opportunities, and courses



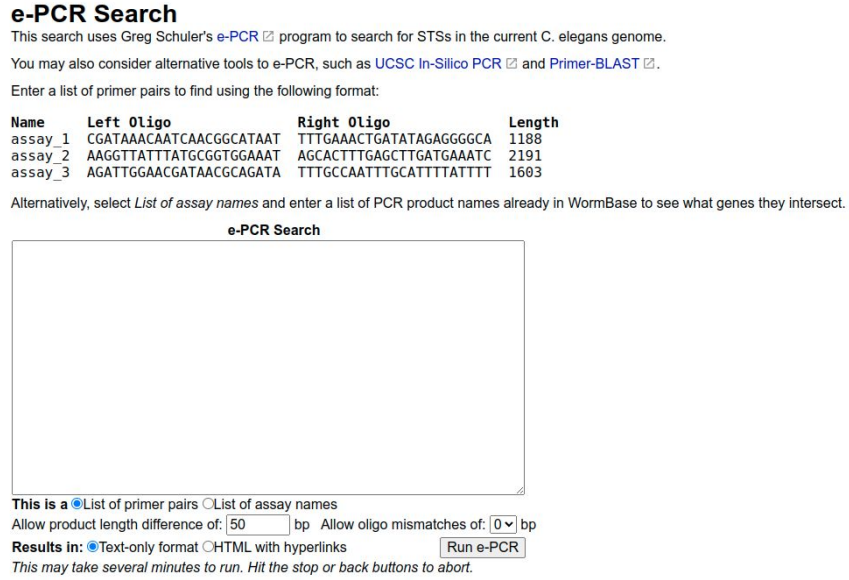
https://flybase.org/
https://doi.org/10.1093/nar/26.1.85

UNIVERSITÄT BONN

WormBase

- Website of the *Cenorhabditis elegans* community

- Tools and resources are similar to FlyBase:
  - Option to run BLAST
  - Check of primers (e-PCR)
  - References to publications
  - Community news



e-PCR Search

This search uses Greg Schuler's e-PCR program to search for STSs in the current C. elegans genome.

You may also consider alternative tools to e-PCR, such as UCSC In-Silico PCR and Primer-BLAST.

Enter a list of primer pairs to find using the following format:

Name	Left Oligo	Right Oligo	Length
assay_1	CGATAAACAATCAACGGCATAAT	TTTGAAACTGATATAGAGGGGCA	1188
assay_2	AAGGTTATTTATGCGGTGGAAAT	AGCACTTTGAGCTTGATGAAATC	2191
assay_3	AGATTGGAACGATAACGCAGATA	TTTGCCAATTTGCATTTTATTTT	1603

Alternatively, select *List of assay names* and enter a list of PCR product names already in WormBase to see what genes they intersect.

e-PCR Search

This is a ◉List of primer pairs ○List of assay names

Allow product length difference of: 50 bp Allow oligo mismatches of: 0 bp

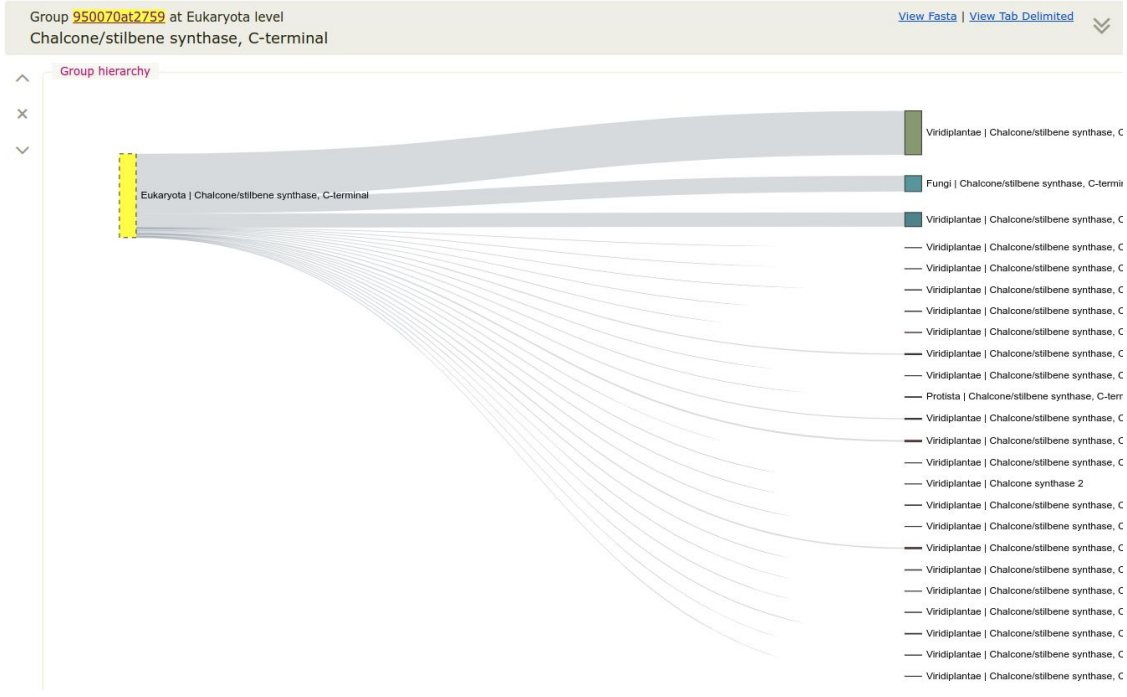
Results in: ◉Text-only format ○HTML with hyperlinks Run e-PCR

This may take several minutes to run. Hit the stop or back buttons to abort.

https://wormbase.org//#012-34-5
Harris et al., 2010: 10.1093/nar/gkp952

UNIVERSITÄT BONN OrthoDB

- Database of similar genes in other species



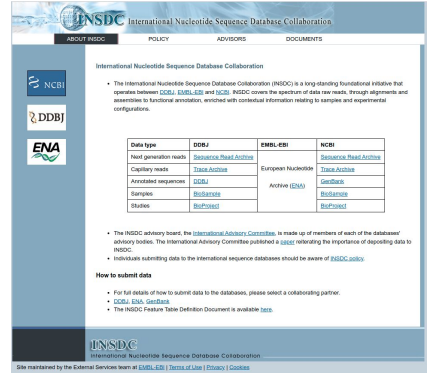
https://www.orthodb.org/
Kriventseva et al., 2018: 10.1093/nar/gky1053

INSDC - a collaborative effort

- INSDC = International Nucleotide Sequence Database Collaboration

- Three databases are mirrored i.e. constantly updated
  - NCBI = National Center for Biotechnology Information (USA)
  - ENA = European Nucleotide Archive (Europe)
  - DDBJ = DNA DataBase of Japan (Japan)

- Users can search in any of the three databases and get the same results



https://www.insdc.org/

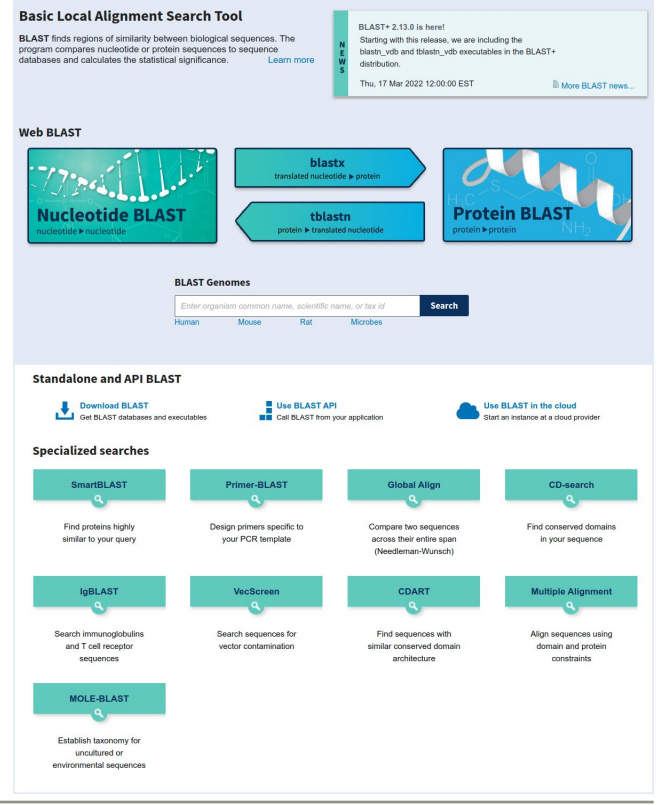
NCBI databases

- NCBI operates a large number of databases and services
- Famous service: BLAST
- Important databases:
  - PubMed: scientific publications
  - GenBank: individual sequences
  - Gene Expression Omnibus (GEO): gene expression data sets
  - Sequence Read Archive (SRA): sequencing data

https://www.ncbi.nlm.nih.gov/

NCBI - BLAST

- BLAST = Basic Local Alignment Search Tool
- Probably the most famous website of the NCBI
- Comparison of sequences against a large database
- Numerous variants of the initial BLASTn were developed

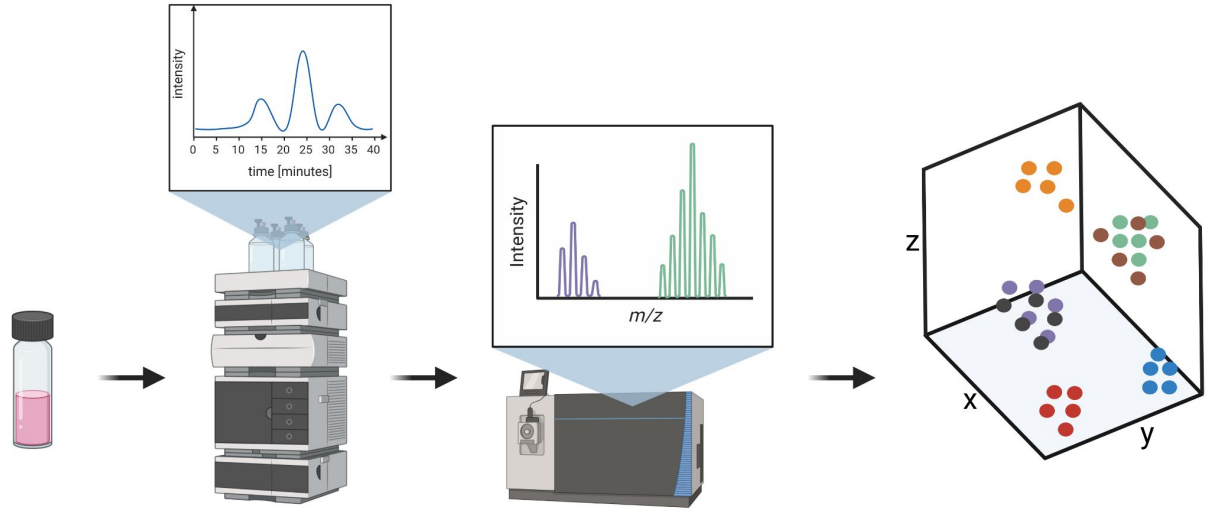


https://blast.ncbi.nlm.nih.gov/Blast.cgi
Altschul et al., 1990: https://doi.org/10.1016/s0022-2836(05)80360-2

Identification of metabolites

- Comparison of HPLC/GC-MS results

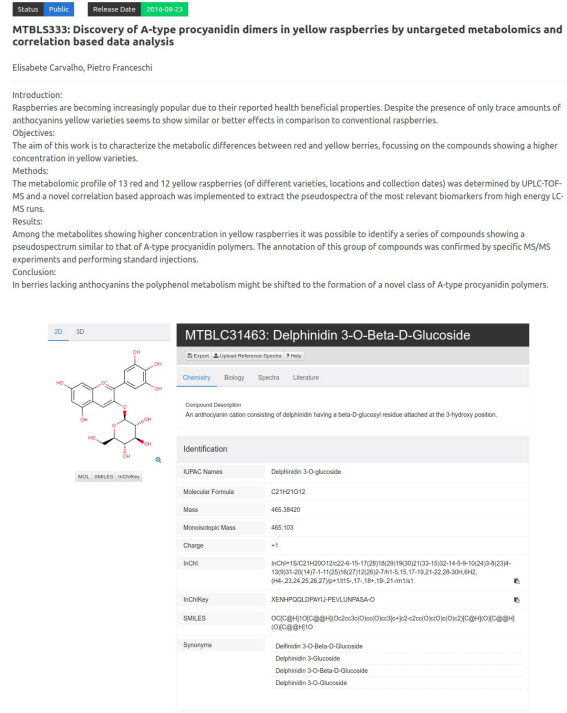
- Authentic standards required for identification of metabolite identity



UNIVERSITÄT BONN

MetaboLights

- Search for studies based on metabolites or species
  - Example: find all studies associated with tyrosine
- Search for information about individual compounds
  - Delphinidin 3-O-β-D-glucoside
- Data sets are available for download



Status Public Release Date 2016-08-23

MTBLS333: Discovery of A-type procyanidin dimers in yellow raspberries by untargeted metabolomics and correlation based data analysis

Elisabete Carvalho, Pietro Franceschi

Introduction:
Raspberries are becoming increasingly popular due to their reported health beneficial properties. Despite the presence of only trace amounts of anthocyanins yellow varieties seems to show similar or better effects in comparison to conventional raspberries.
Objectives:
The aim of this work is to characterize the metabolic differences between red and yellow berries, focussing on the compounds showing a higher concentration in yellow varieties.
Methods:
The metabolomic profile of 13 red and 12 yellow raspberries (of different varieties, locations and collection dates) was determined by UPLC-TOF-MS and a novel correlation based approach was implemented to extract the pseudospectra of the most relevant biomarkers from high energy LC-MS runs.
Results:
Among the metabolites showing higher concentration in yellow raspberries it was possible to identify a series of compounds showing a pseudospectrum similar to that of A-type procyanidin polymers. The annotation of this group of compounds was confirmed by specific MS/MS experiments and performing standard injections.
Conclusion:
In berries lacking anthocyanins the polyphenol metabolism might be shifted to the formation of a novel class of A-type procyanidin polymers.

2D 3D

MTBLC31463: Delphinidin 3-O-Beta-D-Glucoside

Chemistry Biology Spectra Literature

Compound Description
An anthocyanin cation consisting of delphinidin having a beta-D-glucosyl residue attached at the 3-hydroxy position.

Identification

IUPAC Names	Delphinidin 3-O-glucoside
Molecular Formula	C21H21O12
Mass	465.38420
Monoisotopic Mass	465.103
Charge	+1
InChI	InChI=1S/C21H20O12/c22-6-15-17(26)18(29)19(30)21(33-15)32-14-5-9-10(24)3-8(23)4-13(9)31-20(14)7-1-11(25)16(27)12(28)2-7/h1-5,15,17-19,21-22,26-30H,6H2,(H4-,23,24,25,26,27)/p+1/t15-,17-,18+,19-,21-/m1/s1
InChIKey	XENHPQQLDPAYIJ-PEVLUNPASA-O
SMILES	OC[C@H]1O[C@@H](Oc2cc3c(O)cc(O)cc3[o+]c2-c2cc(O)c(O)c(O)c2)[C@H](O)[C@@H](O)[C@@H]1O
Synonyms	Delfinidin 3-O-Beta-D-Glucoside Delphinidin 3-Glucoside Delphinidin 3-O-Beta-D-Glucoside Delphinidin 3-O-Glucoside

Download data from MetaboLights

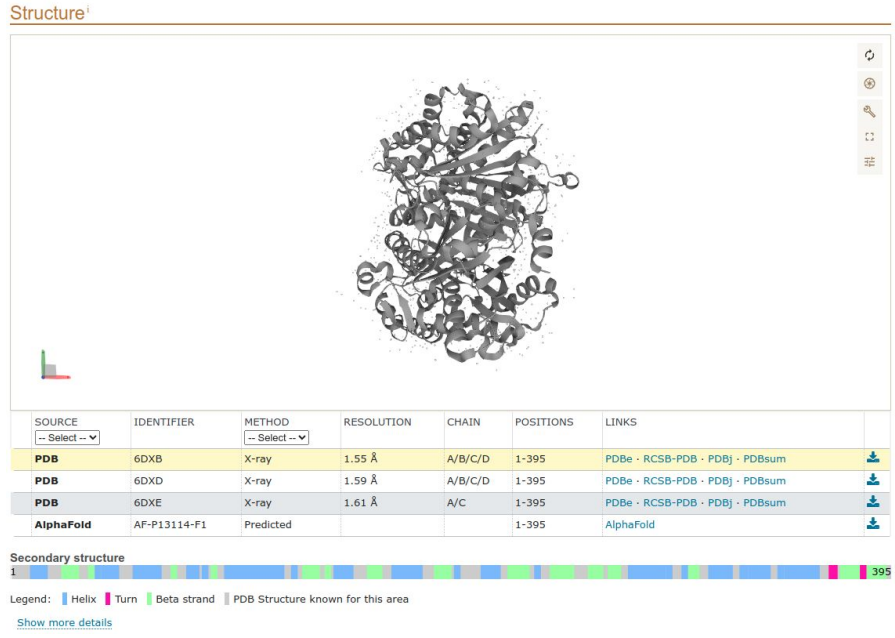
MetaboLights ISAcreator software download. To make it easy for new users, please download and just unzip our pre-packaged ISAcreator with plugin and configurations.

Experiments. All public MetaboLights experiments can be downloaded from our public ftp archive. Please find zip archives under the "studies" folder. Each public study can be found in the corresponding MTBLS-id folder. Complete experiments can be opened with ISAcreator or you can extract the archives using your normal unzip program.

https://www.ebi.ac.uk/metabolights/

Identification of enzymes - UniProt/SwissProt

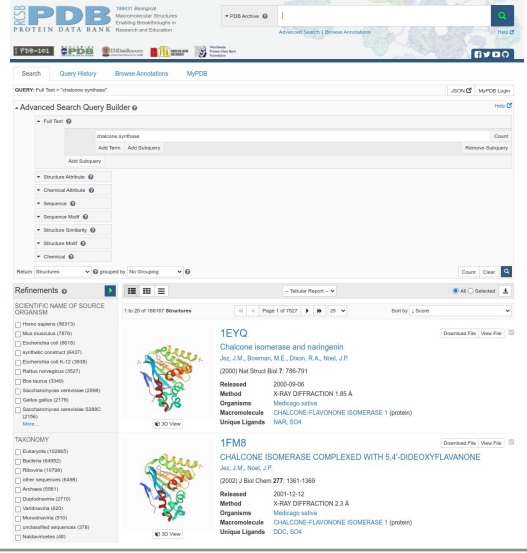
- Curated database of protein sequences
- Provides additional information retrieved from other databases
- 2D and 3D structure of protein
- Sequence available for download

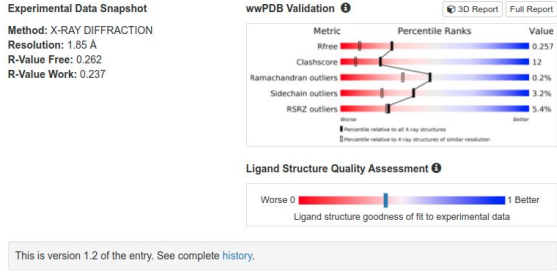


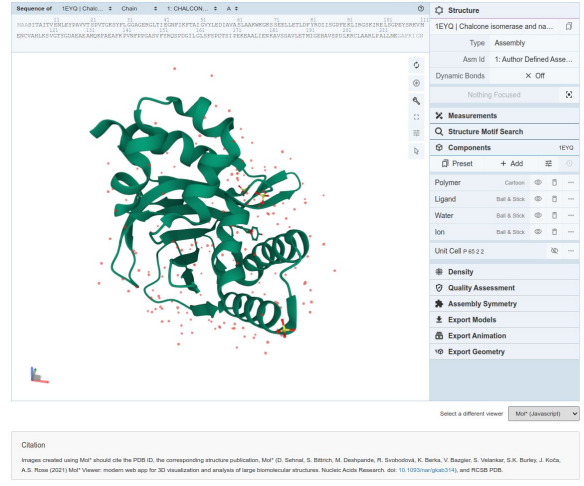
https://www.uniprot.org/

PDB: Protein Data Bank

- Large collection of protein structures
- Download of protein structure files possible
- PDB viewer (DeepView) allows interactive inspection of protein structures





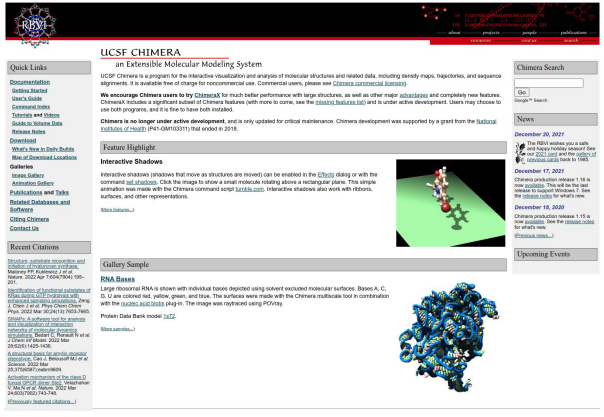


https://www.rcsb.org/; https://spdbv.unil.ch/main_tut.html
Zardecki et al., 2022: 10.1002/pro.4200

- Downloaded PDB files can be displayed in external tools

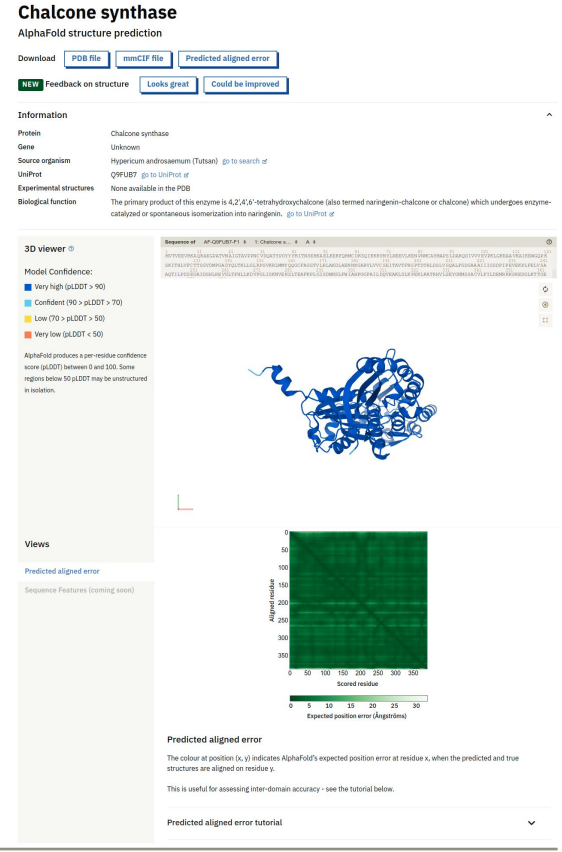
- Chimera: https://www.cgl.ucsf.edu/chimera/

- PyMol: https://pymol.org/2/



AlphaFold DB: Protein Structure Database

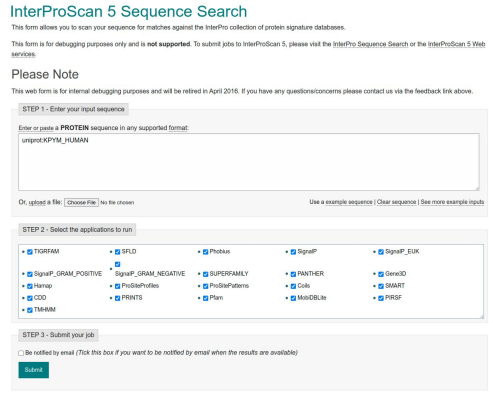
- AlphaFold is an Artificial Intelligence tool developed by DeepMind
  - Leading tool in the prediction of protein structures
- Corresponding database is maintained by EMBL-EBI
- Database contains predicted protein structures



https://alphafold.ebi.ac.uk/
Jumper et al., 2021: 10.1038/s41586-021-03819-2

Identification of enzymes - InterProScan5

- Screen of protein sequences against collection of protein signatures
- Allows the assignment of functional annotation terms
- Available as web service, but also as stand alone tool



Results for job iprscan5-I20220320-102557-0980-87120812-p2m

Tool Output | Submission Details

Download in XML format | Download in TSV format | Download in GFF3 format | Download in SVG format | Download HTML tarball file | Download in JSON format

sp|P13114|CHSY_ARATH	ba46b0f06bfbf1c91d161f0f91310f43	395	Gene3D	G3DSA:3.40.47.10	-	1	241	4.8E-101	T	20-03-2022	
sp|P13114|CHSY_ARATH	ba46b0f06bfbf1c91d161f0f91310f43	395	SUPERFAMILY	SSF53901	Thiolase-like	241	393	2.98E-51	T	20-03-20	
sp|P13114|CHSY_ARATH	ba46b0f06bfbf1c91d161f0f91310f43	395	SUPERFAMILY	SSF53901	Thiolase-like	10	237	2.38E-78	T	20-03-20	
sp|P13114|CHSY_ARATH	ba46b0f06bfbf1c91d161f0f91310f43	395	Pfam	PF02797	Chalcone and stilbene synthases, C-terminal domain	244	394	1.5E-71			
sp|P13114|CHSY_ARATH	ba46b0f06bfbf1c91d161f0f91310f43	395	PANTHER	PTHR11877:SF81	BNAA02G30320D PROTEIN	6	394	0.0	T	20-03-2022	
sp|P13114|CHSY_ARATH	ba46b0f06bfbf1c91d161f0f91310f43	395	Gene3D	G3DSA:3.40.47.10	-	242	395	9.7E-62	T	20-03-2022	IPR01603
sp|P13114|CHSY_ARATH	ba46b0f06bfbf1c91d161f0f91310f43	395	PANTHER	PTHR11877	HYDROXYMETHYLGLUTARYL-COA SYNTHASE	6	394	0.0	T		
sp|P13114|CHSY_ARATH	ba46b0f06bfbf1c91d161f0f91310f43	395	CDD	cd00831 CHS_like	21	390	0.0	T	20-03-2022	-	-
sp|P13114|CHSY_ARATH	ba46b0f06bfbf1c91d161f0f91310f43	395	PIRSF	PIRSF000451	PKS_III 7	394	0.0	T	20-03-2022	IPR011141	
sp|P13114|CHSY_ARATH	ba46b0f06bfbf1c91d161f0f91310f43	395	Pfam	PF00195	Chalcone and stilbene synthases, N-terminal domain	10	233	2.9E-119			
sp|P13114|CHSY_ARATH	ba46b0f06bfbf1c91d161f0f91310f43	395	ProSitePatterns	PS00441	Chalcone and stilbene synthases active site.	161	177	-			

http://www.ebi.ac.uk/Tools/services/web/toolform.ebi?tool=iprscan5&sequence=uniprot:KPYM_HUMAN

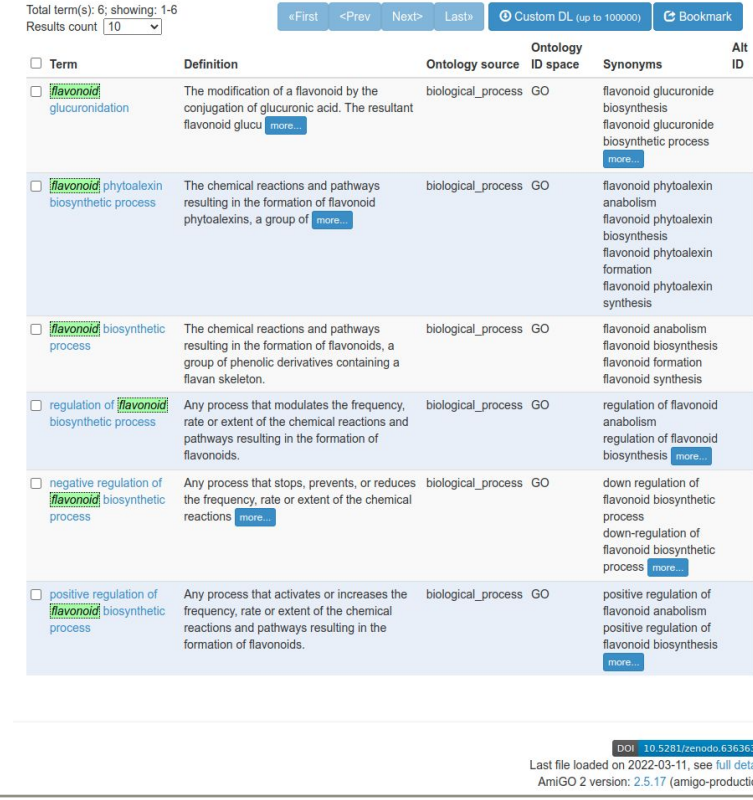
Jones et al., 2014: 10.1093/bioinformatics/btu031

Gene Ontology (GO) and AmiGO

- Defined statements about the function of a gene (controlled vocabulary)
- Hierarchical structure
  - Example: ‘metabolic process’ > ‘biosynthetic process’ > … > ‘chalcone synthase’
- Supported by the Alliance of Genome Resources
- Connected with various other databases e.g. TAIR, FlyBase, Reactome, UniProt
- Machine readable to allow automatic processing
- Tools: Blast2GO and AmiGO
  - analyze the function of a sequence (web service and standalone)

http://amigo.geneontology.org/amigo/term/GO:0009813
The Gene Ontology Consortium, 2007: 10.1093/nar/gkm883

- AmiGO allows text or sequence searches
- Options to explore annotations in detail
- Database covers various kingdoms (not restricted to plants)



Total term(s): 6; showing: 1-6
Results count 10

«First | <Prev | Next> | Last» | Custom DL (up to 100000) | Bookmark

Term	Definition	Ontology source	Ontology ID space	Synonyms	Alt ID
flavonoid glucuronidation	The modification of a flavonoid by the conjugation of glucuronic acid. The resultant flavonoid glucu more...	biological_process	GO	flavonoid glucuronide biosynthesis flavonoid glucuronide biosynthetic process more...	
flavonoid phytoalexin biosynthetic process	The chemical reactions and pathways resulting in the formation of flavonoid phytoalexins, a group of more...	biological_process	GO	flavonoid phytoalexin anabolism flavonoid phytoalexin biosynthesis flavonoid phytoalexin formation flavonoid phytoalexin synthesis	
flavonoid biosynthetic process	The chemical reactions and pathways resulting in the formation of flavonoids, a group of phenolic derivatives containing a flavan skeleton.	biological_process	GO	flavonoid anabolism flavonoid biosynthesis flavonoid formation flavonoid synthesis	
regulation of flavonoid biosynthetic process	Any process that modulates the frequency, rate or extent of the chemical reactions and pathways resulting in the formation of flavonoids.	biological_process	GO	regulation of flavonoid anabolism regulation of flavonoid biosynthesis more...	
negative regulation of flavonoid biosynthetic process	Any process that stops, prevents, or reduces the frequency, rate or extent of the chemical reactions more...	biological_process	GO	down regulation of flavonoid biosynthetic process down-regulation of flavonoid biosynthetic process more...	
positive regulation of flavonoid biosynthetic process	Any process that activates or increases the frequency, rate or extent of the chemical reactions and pathways resulting in the formation of flavonoids.	biological_process	GO	positive regulation of flavonoid anabolism positive regulation of flavonoid biosynthesis more...	

DOI 10.5281/zenodo.6363634
Last file loaded on 2022-03-11, see full details
AmiGO 2 version: 2.5.17 (amigo-production)

http://amigo.geneontology.org/amigo/
Carbon et al., 2009: 10.1093/bioinformatics/btn615

Pfam: Protein family database

- Assignment of protein functions based on Hidden Markov Models (HMMs)

- Sequences are screened based on HMM profile

Sequences are aligned

↓

Modelling of insertions/deletions

↓

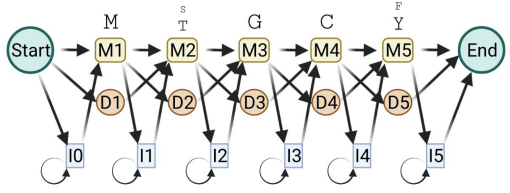
Amino acid frequency at each alignment position is encoded

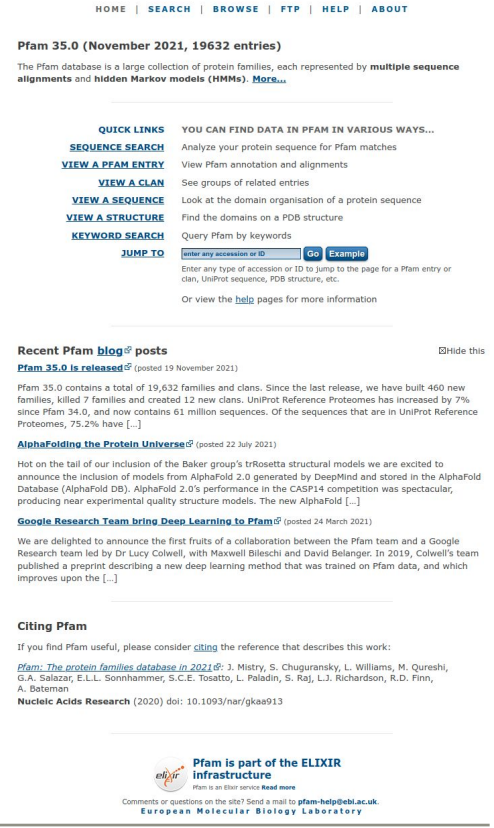
↓

HMM profile generated

```
seq1    MTGC-Y
seq2    MSGC-F
seq3    MTGC-Y
seq4    M-GCAY
        123456
```

2 = deletion
5 = insertion




HOME | SEARCH | BROWSE | FTP | HELP | ABOUT

Pfam 35.0 (November 2021, 19632 entries)

The Pfam database is a large collection of protein families, each represented by **multiple sequence alignments** and **hidden Markov models (HMMs)**. More...

QUICK LINKS	YOU CAN FIND DATA IN PFAM IN VARIOUS WAYS...
SEQUENCE SEARCH	Analyze your protein sequence for Pfam matches
VIEW A PFAM ENTRY	View Pfam annotation and alignments
VIEW A CLAN	See groups of related entries
VIEW A SEQUENCE	Look at the domain organisation of a protein sequence
VIEW A STRUCTURE	Find the domains on a PDB structure
KEYWORD SEARCH	Query Pfam by keywords
JUMP TO	enter any accession or ID · Go · Example

Enter any type of accession or ID to jump to the page for a Pfam entry or clan, UniProt sequence, PDB structure, etc.

Or view the help pages for more information

Recent Pfam blog posts ☒Hide this

Pfam 35.0 is released (posted 19 November 2021)

Pfam 35.0 contains a total of 19,632 families and clans. Since the last release, we have built 460 new families, killed 7 families and created 12 new clans. UniProt Reference Proteomes has increased by 7% since Pfam 34.0, and now contains 61 million sequences. Of the sequences that are in UniProt Reference Proteomes, 75.2% have [...]

AlphaFolding the Protein Universe (posted 22 July 2021)

Hot on the tail of our inclusion of the Baker group's trRosetta structural models we are excited to announce the inclusion of models from AlphaFold 2.0 generated by DeepMind and stored in the AlphaFold Database (AlphaFold DB). AlphaFold 2.0's performance in the CASP14 competition was spectacular, producing near experimental quality structure models. The new AlphaFold [...]

Google Research Team bring Deep Learning to Pfam (posted 24 March 2021)

We are delighted to announce the first fruits of a collaboration between the Pfam team and a Google Research team led by Dr Lucy Colwell, with Maxwell Bileschi and David Belanger. In 2019, Colwell's team published a preprint describing a new deep learning method that was trained on Pfam data, and which improves upon the [...]

Citing Pfam

If you find Pfam useful, please consider citing the reference that describes this work:

Pfam: The protein families database in 2021: J. Mistry, S. Chuguransky, L. Williams, M. Qureshi, G.A. Salazar, E.L.L. Sonnhammer, S.C.E. Tosatto, L. Paladin, S. Raj, L.J. Richardson, R.D. Finn, A. Bateman
Nucleic Acids Research (2020) doi: 10.1093/nar/gkaa913

Pfam is part of the ELIXIR infrastructure
Pfam is an Elixir service Read more

Comments or questions on the site? Send a mail to pfam-help@ebi.ac.uk.
European Molecular Biology Laboratory

http://pfam.xfam.org/
Mistry et al., 2020: 10.1093/nar/gkaa913

UNIVERSITÄT BONN

Mercator: functional annotation of protein sequences

- Online tool for the annotation of all protein sequences in a submitted FASTA file

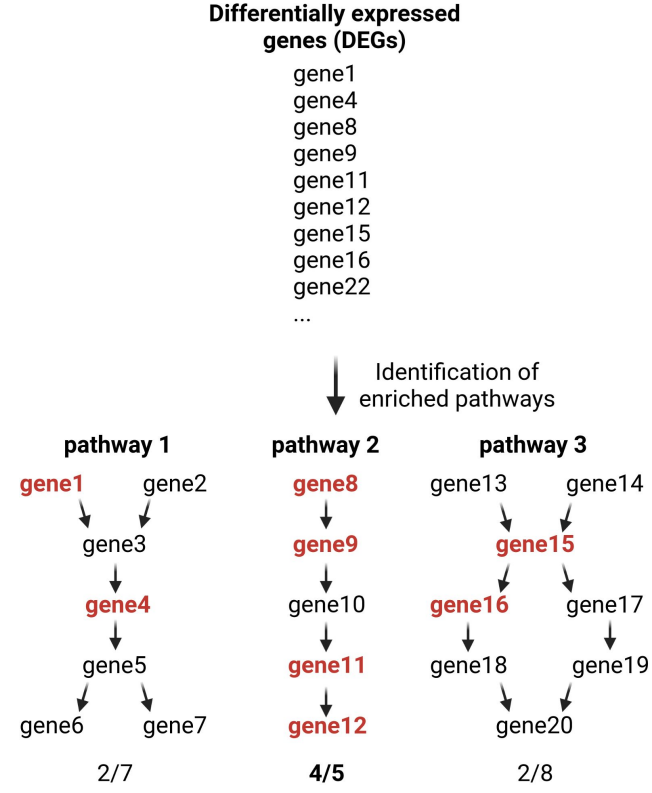
- FASTA format:
  - Header line starts with '>' followed by the sequence name
  - Header is followed by unrestricted number of sequence lines

```
>TRINITY_DN100013_c0_g1_i1
MIGPMPGMEGKLMLPGASPVGLEVVLSLVVASTPILSDTSLCTPSFYHFLLLLGPITSISNLIVRPFFLTLLSSITVIYGRLCFFAFDFYVY
>TRINITY_DN100016_c0_g1_i1
MPKKSSNIQQNNQNRIKRKGRPPKKKKYVQQIDSSDEDILSVRHASTRPRIISIRRNEIPMRPEIHIPPTSIPSFNNPNNIHQSSSDRQMPP
>TRINITY_DN100019_c0_g1_i1
MPPKAADKKPAAKAPATASKAPEKKDAGKKTAATGDKKKRTKARKETYSSYIYKVLKQVHPDTGISNRAMSILNSFVNDIFERVATEASKLA
>TRINITY_DN10001_c0_g1_i2
MAKVGNPIVDIETDGSVNEPESSEKNIEVSSSSTQAPESTNTTELLVNEKKAFSLATPAVRRVAREHNIDINNIKGTGKNGRITKEDILNYV
>TRINITY_DN100025_c0_g1_i1
MVENQDGCFFKPGWKEFVRSNDLEGGDFLVFNLVDKISYQVVIFDGTCCPKDLCPFSIMNPIFQHLRNKIFLSKKEEIKLKGNRKVHSVNEN
>TRINITY_DN10002_c0_g1_i1
MSDELNQVFQRCREQKRPVFVAFVTAGYPDSEETVDILLGLEAGGADIIELGIPFTDPMVDGKTIQDANNVALENKIDISKCLSYVSESRAK
```


https://plabipd.de/portal/mercator-sequence-annotation
Lohse et al., 2014: 10.1111/pce.12231; Haak et al., 2018: 10.3389/fmolb.2018.00062

Concept of enrichment analyses

- Omics analyses can identify a large set of differentially expressed genes
- Making sense of large gene sets is challenging
- Identification of a common function is necessary to reduce complexity
- Enrichment analyses can help to identify shared functions or pathways



Enrichment analyses - GO

- Functional enrichment analyses can be based on GO terms
- Assignment of GO terms to genes of the species of interest
- Tools for GO enrichment analysis:
  - PANTHER: http://geneontology.org/
  - g:Profiler: http://biit.cs.ut.ee/gprofiler/gost
  - GOrilla: http://cbl-gorilla.cs.technion.ac.il/

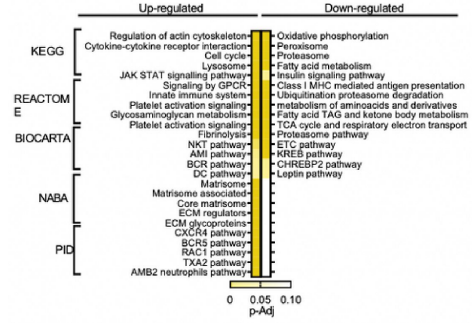
Enrichment analyses - KEGG

- Identification of enriched KEGG pathways among a gene set

- Available tools:
  - KEGG Mapper: https://www.genome.jp/kegg/kegg1b.html
  - pathfindeR: https://doi.org/10.3389/fgene.2019.00858

- Names of pathways is not always informative

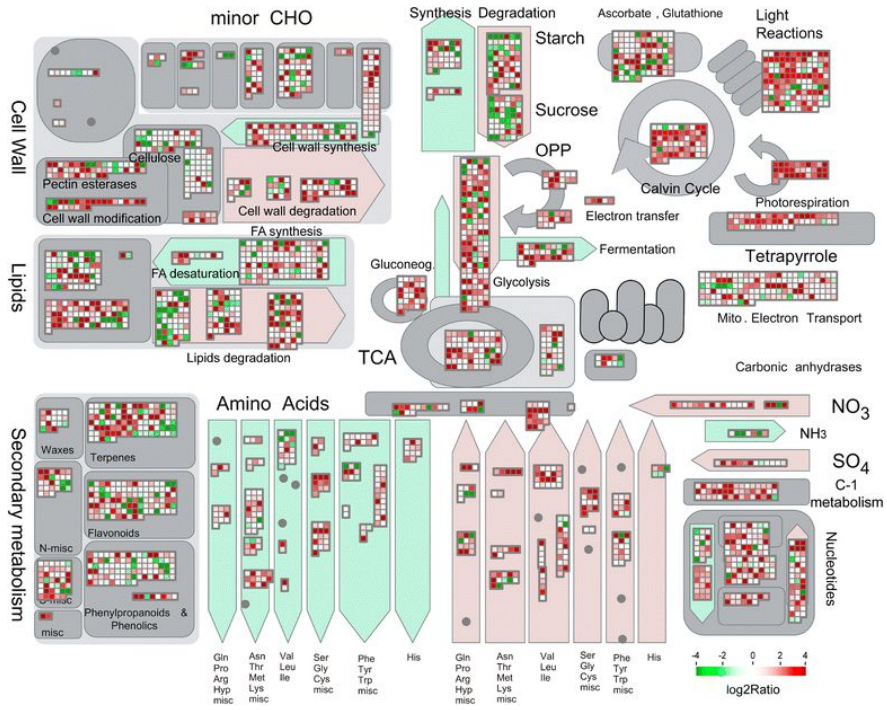
- KEGG pathways can be too superficial and are only a first step in analyses



https://doi.org/10.21203/rs.3.rs-57182/v1

Enrichment analyses - MapMan

- Identification of enriched pathways
- Visualization of up- and down-regulated pathways:
  - Red = up-regulation
  - Green = down-regulation
  - White = neutral (no change)
- Overview of the metabolism of a cell
- MapMan: https://mapman.gabipd.org/mapman



Gao et al., 2015: 10.1186/s12864-015-1633-y

Summary

- Types of databases (reactions, enzymes, sequences, metabolites, species-specific)
- Species-specific databases
- Finding information about enzymes, transcription factors, and other genes
- Finding information about protein structures
- NCBI & INSDC
- Enrichment analyses

UNIVERSITÄT BONN

Time for questions!

UNIVERSITÄT BONN

Questions

1. What is the meaning of FAIR data?
2. How would you get insights into the function of an unknown sequence?
3. How can you identify the enzymes involved in the flavonol biosynthesis?
4. Where can you find information about the substrate affinity of the chalcone synthase?
5. What is TAIR?
6. Where can you find information about the expression of *MYB12* in different plant organs?
7. Where can you get the genome sequence and corresponding annotation of *Marchantia polymorpha*?
8. Which databases are connected through INSDC?
9. Where can you find the results of metabolomic studies?
10. Where can you find information about the structure of CHS?
11. Which tools can be applied to annotate large sets of (poly)peptide sequences?
12. Which tools can be applied for pathway enrichment analyses?
13. What are the two line types in a FASTA file?