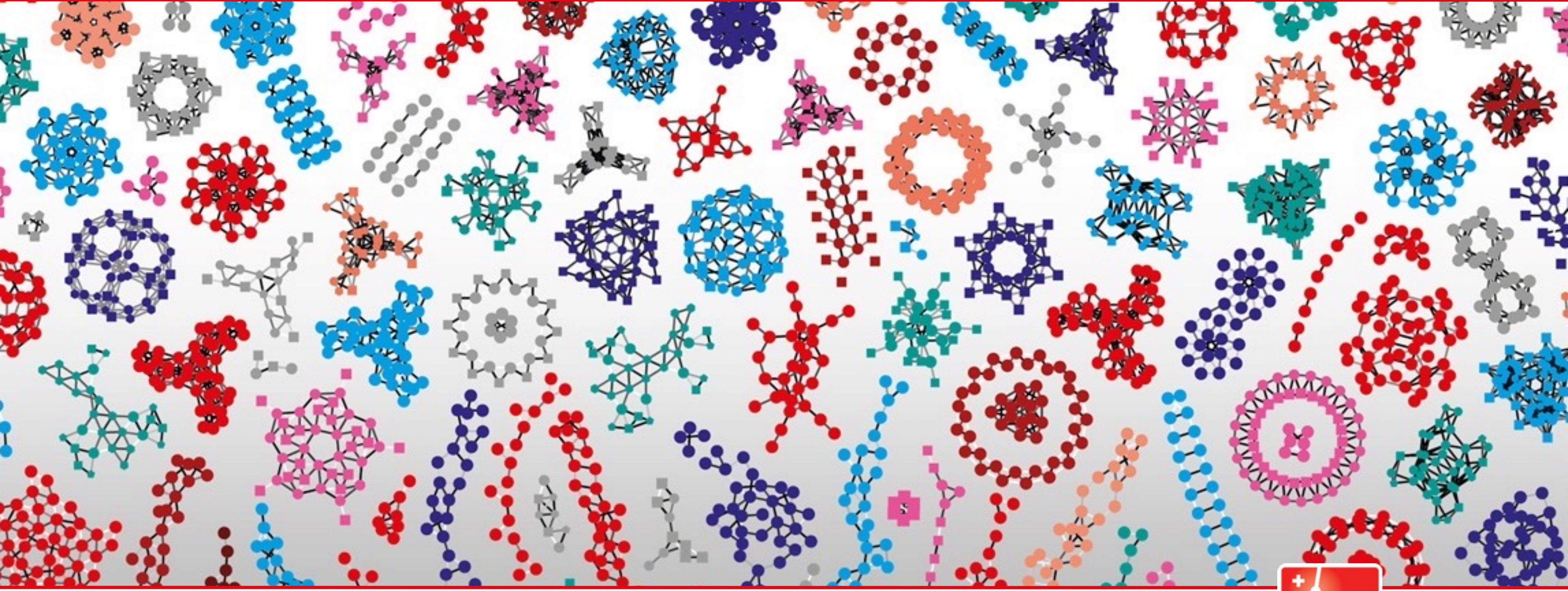
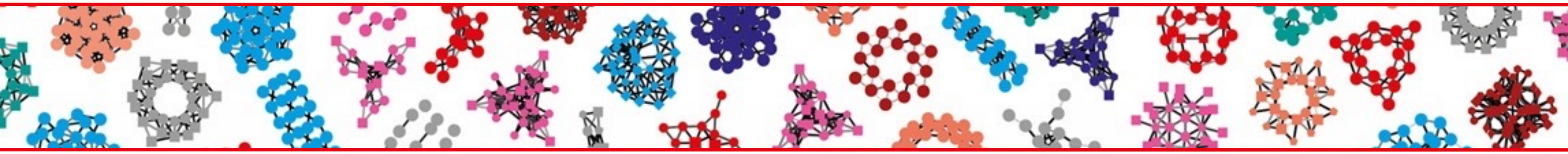


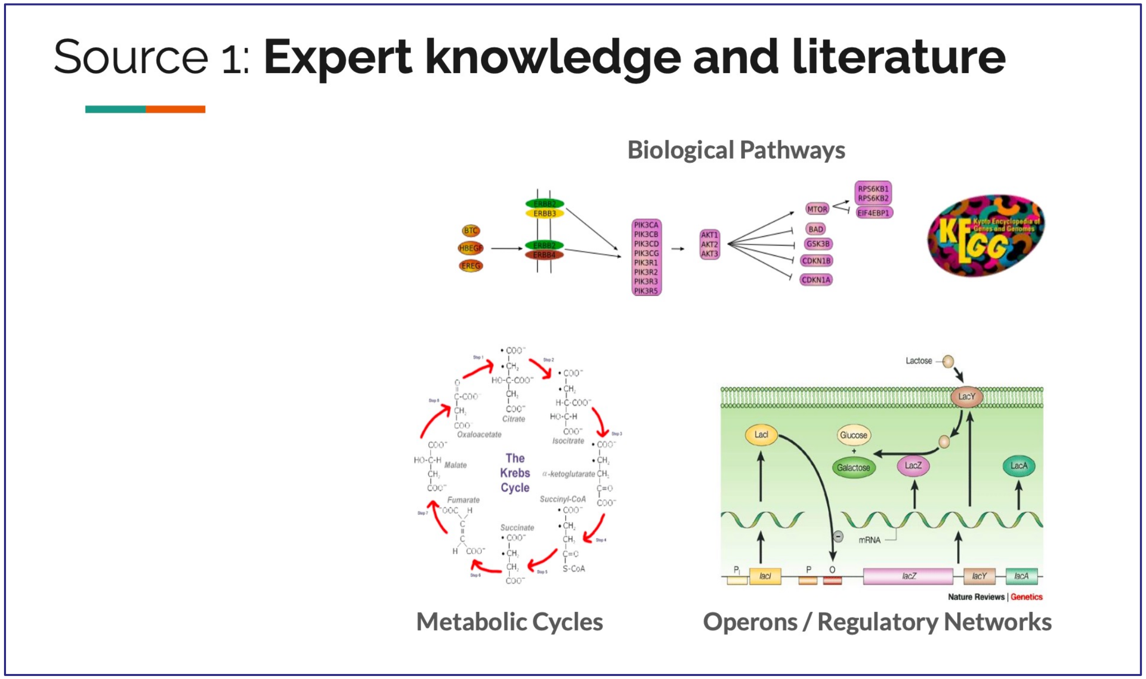
- Identifiers, cross-references and graphs
- MetaNetX
- Diffusion on graphs

Summer School in Aussois
6th of September 2023
Marco Pagni

www.sib.swiss



Database structures



Source 1: **Expert knowledge and literature**

Biological Pathways

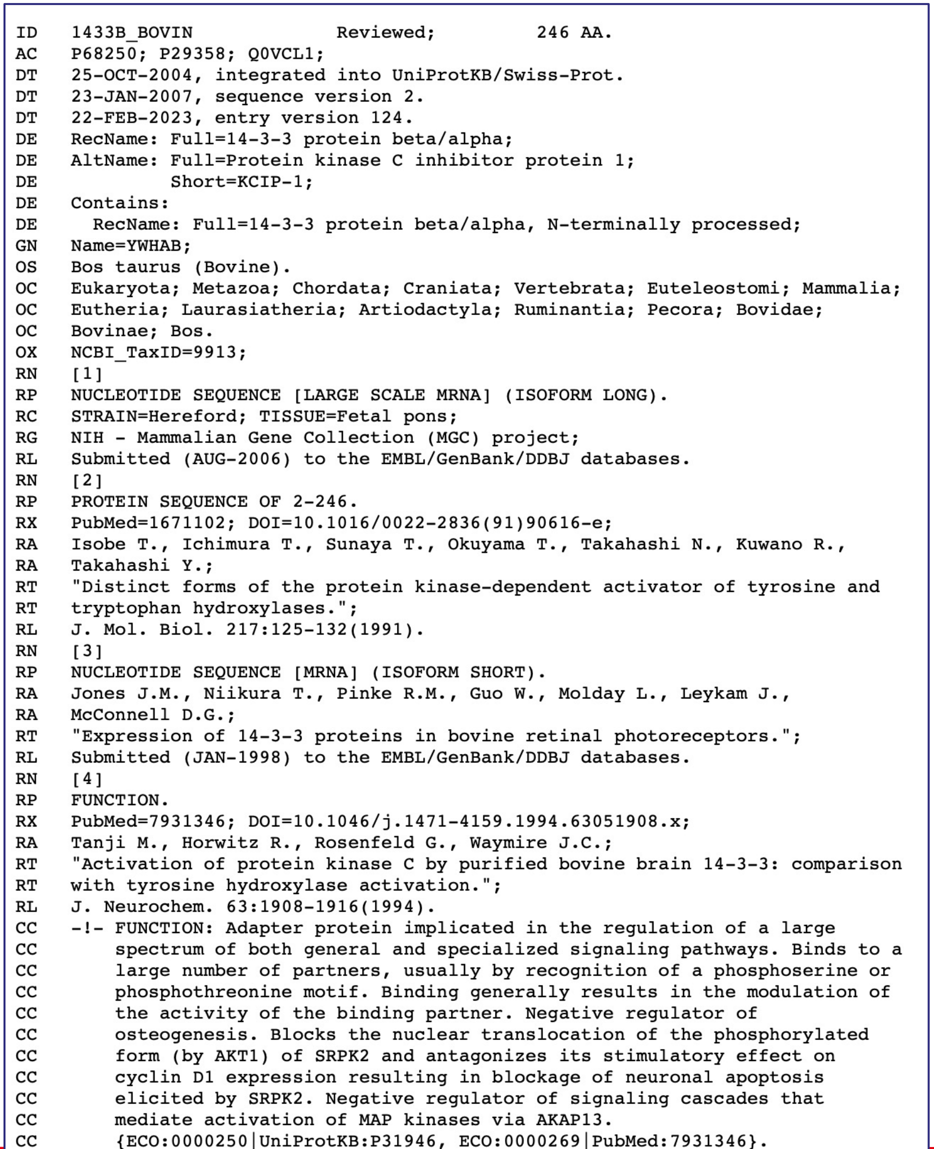
Metabolic Cycles

Operons / Regulatory Networks

figure from Galadriel Briere presentation on Tuesday

A look back in time

- Database best practices have not really changed over the last 25 years
- User interfaces have dramatically improved



```
ID   1433B_BOVIN             Reviewed;         246 AA.
AC   P68250; P29358; Q0VCL1;
DT   25-OCT-2004, integrated into UniProtKB/Swiss-Prot.
DT   23-JAN-2007, sequence version 2.
DT   22-FEB-2023, entry version 124.
DE   RecName: Full=14-3-3 protein beta/alpha;
DE   AltName: Full=Protein kinase C inhibitor protein 1;
DE            Short=KCIP-1;
DE   Contains:
DE     RecName: Full=14-3-3 protein beta/alpha, N-terminally processed;
GN   Name=YWHAB;
OS   Bos taurus (Bovine).
OC   Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia;
OC   Eutheria; Laurasiatheria; Artiodactyla; Ruminantia; Pecora; Bovidae;
OC   Bovinae; Bos.
OX   NCBI_TaxID=9913;
RN   [1]
RP   NUCLEOTIDE SEQUENCE [LARGE SCALE MRNA] (ISOFORM LONG).
RC   STRAIN=Hereford; TISSUE=Fetal pons;
RG   NIH - Mammalian Gene Collection (MGC) project;
RL   Submitted (AUG-2006) to the EMBL/GenBank/DDBJ databases.
RN   [2]
RP   PROTEIN SEQUENCE OF 2-246.
RX   PubMed=1671102; DOI=10.1016/0022-2836(91)90616-e;
RA   Isobe T., Ichimura T., Sunaya T., Okuyama T., Takahashi N., Kuwano R.,
RA   Takahashi Y.;
RT   "Distinct forms of the protein kinase-dependent activator of tyrosine and
RT   tryptophan hydroxylases.";
RL   J. Mol. Biol. 217:125-132(1991).
RN   [3]
RP   NUCLEOTIDE SEQUENCE [MRNA] (ISOFORM SHORT).
RA   Jones J.M., Niikura T., Pinke R.M., Guo W., Molday L., Leykam J.,
RA   McConnell D.G.;
RT   "Expression of 14-3-3 proteins in bovine retinal photoreceptors.";
RL   Submitted (JAN-1998) to the EMBL/GenBank/DDBJ databases.
RN   [4]
RP   FUNCTION.
RX   PubMed=7931346; DOI=10.1046/j.1471-4159.1994.63051908.x;
RA   Tanji M., Horwitz R., Rosenfeld G., Waymire J.C.;
RT   "Activation of protein kinase C by purified bovine brain 14-3-3: comparison
RT   with tyrosine hydroxylase activation.";
RL   J. Neurochem. 63:1908-1916(1994).
CC   -!- FUNCTION: Adapter protein implicated in the regulation of a large
CC       spectrum of both general and specialized signaling pathways. Binds to a
CC       large number of partners, usually by recognition of a phosphoserine or
CC       phosphothreonine motif. Binding generally results in the modulation of
CC       the activity of the binding partner. Negative regulator of
CC       osteogenesis. Blocks the nuclear translocation of the phosphorylated
CC       form (by AKT1) of SRPK2 and antagonizes its stimulatory effect on
CC       cyclin D1 expression resulting in blockage of neuronal apoptosis
CC       elicited by SRPK2. Negative regulator of signaling cascades that
CC       mediate activation of MAP kinases via AKAP13.
CC       {ECO:0000250|UniProtKB:P31946, ECO:0000269|PubMed:7931346}.
```

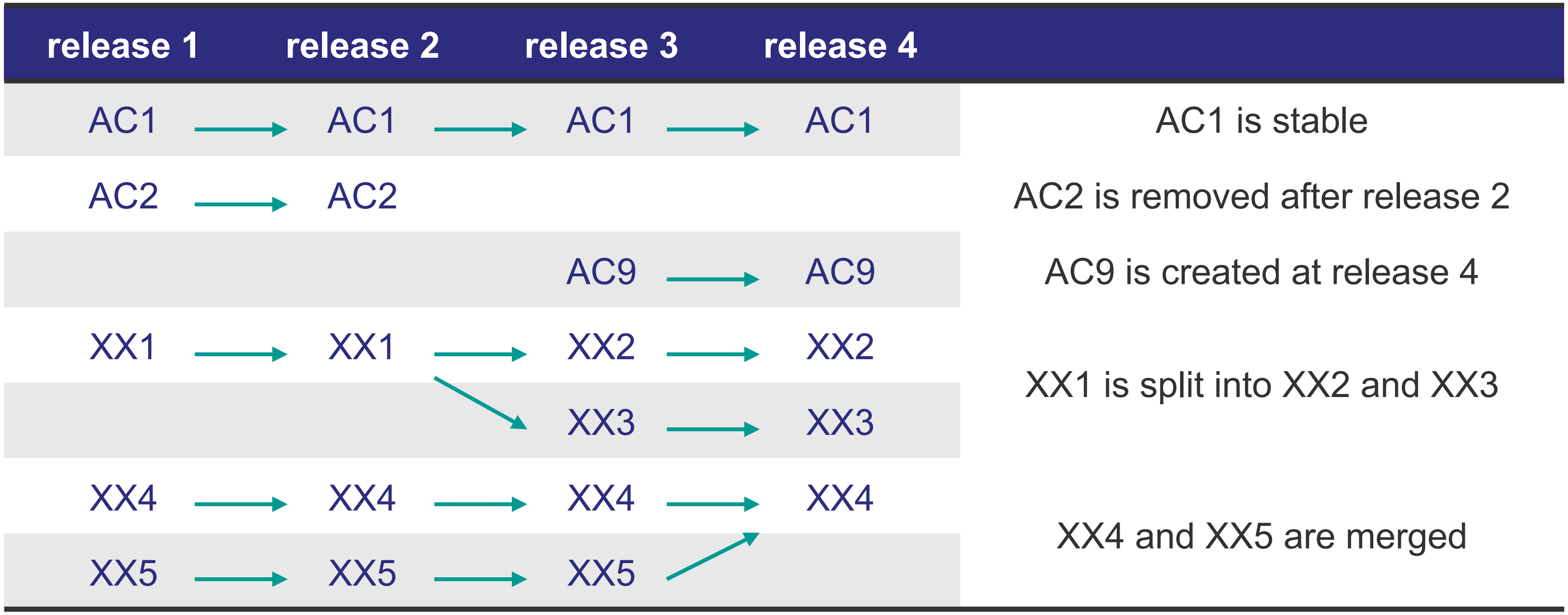
Database entry life cycle

An accession number:

- uniquely identify an "entry" across releases
- is a string, not a number
- is opaque
- is meant to be stable across releases
- must never be recycled

An entry:

- contains a core of information, secondary information and cross-references
- hopefully improves across releases
- typically contains more false negatives than false positives



release 1	release 2	release 3	release 4	
AC1	AC1	AC1	AC1	AC1 is stable
AC2	AC2			AC2 is removed after release 2
		AC9	AC9	AC9 is created at release 4
XX1	XX1	XX2	XX2	XX1 is split into XX2 and XX3
		XX3	XX3	
XX4	XX4	XX4	XX4	XX4 and XX5 are merged
XX5	XX5	XX5		

→ forms a directed acyclic graph (DAG)

Mnemonic identifier

A mnemonic identifier:

- is meant to facilitate human life
- should not be used as a stable reference
- is not necessarily propagated across releases

- mnemonic
- primary accession number
- secondary (deprecated) accession numbers

```
ID   1433B_BOVIN          Reviewed;        246 AA.
AC   P68250; P29358; Q0VCL1;
DT   25-OCT-2004, integrated into UniProtKB/Swiss-Prot.
DT   23-JAN-2007, sequence version 2.
DT   22-FEB-2023, entry version 124.
DE   RecName: Full=14-3-3 protein beta/alpha;
DE   AltName: Full=Protein kinase C inhibitor protein 1;
DE            Short=KCIP-1;
DE   Contains:
DE     RecName: Full=14-3-3 protein beta/alpha, N-terminally
processed;
```

Some resources have no mnemonic identifier. In ChEBI is found an accession number and a molecule name

Some resources do not distinguish accession number and mnemonic identifier. For example, this is found in some metabolic models

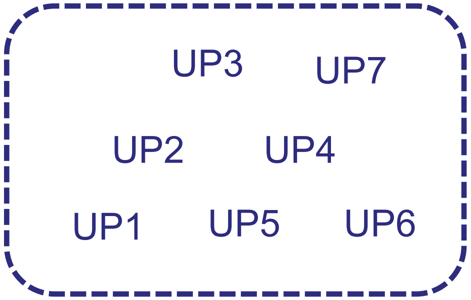
Gene names are rather on the "mnemonic side". ENSEMBL identifiers are accession number linked to a particular genome assembly.

Recommendations: work with mnemonic identifiers when available because they are more informative, but always keep track of the accession numbers.

Database structure

The overall database structure may consist in a set of independent entries. For example

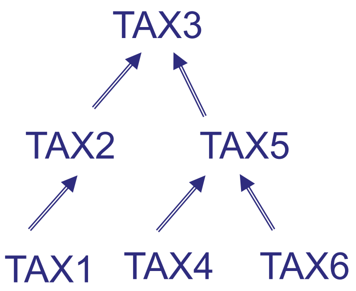
- **UniProt**
- **EMBL**



Database structure

In a tree structure, every entry (node) has zero or one parent entry.

- **The NCBI taxonomy is a single huge tree**
- **Medical Subject Headings (MeSH) are made of 16 trees.**



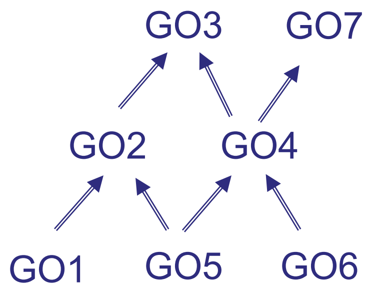
entry	parent
TAX1	TAX2
TAX2	TAX3
TAX3	
TAX4	TAX5
TAX5	TAX3
TAX6	TAX5

Database structure

In a directed acyclic graph (DAG) every entry (node) has zero, one or multiple parent entries.

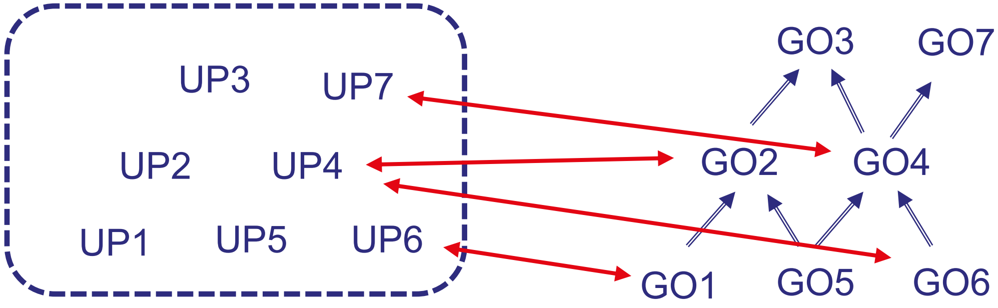
This can be referred to as an ontology, when the properties attributed to parents are inherited by their children

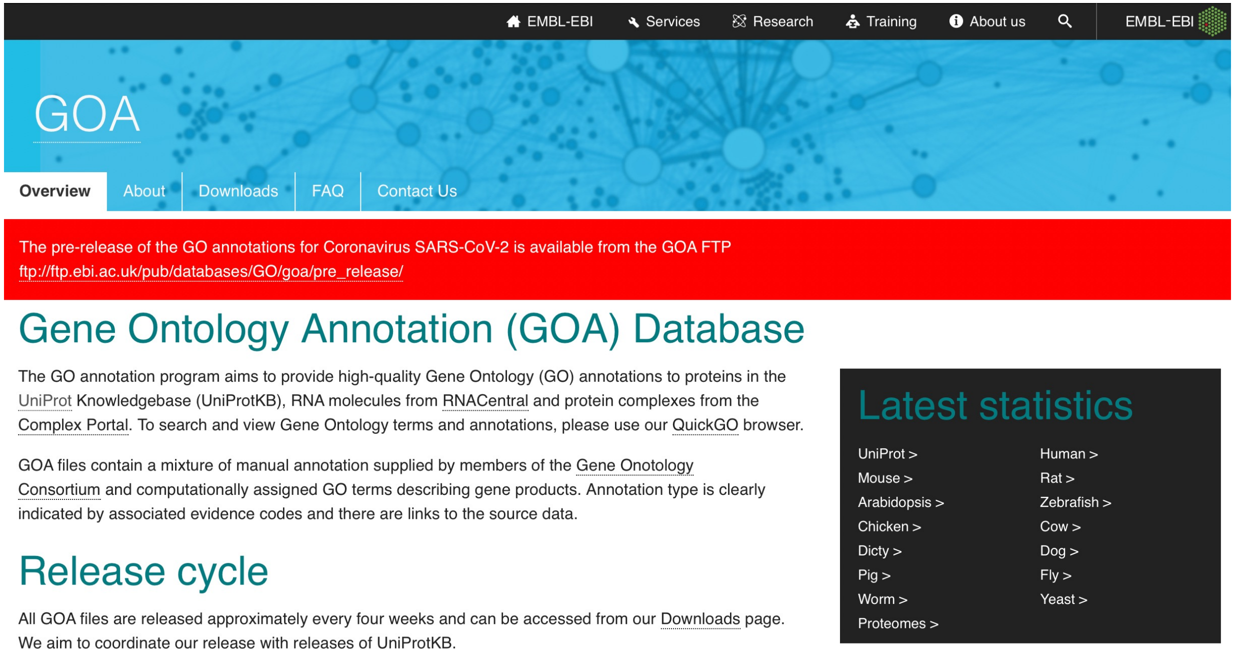
- **GENE Ontology (GO)**
- **ChEBI ontology**



entry	parent
GO1	GO2
GO2	GO3
GO3	
GO4	GO3, GO7
GO5	GO4
GO6	GO4
GO7	

About GO and GOA





The pre-release of the GO annotations for Coronavirus SARS-CoV-2 is available from the GOA FTP
ftp://ftp.ebi.ac.uk/pub/databases/GO/goa/pre_release/

Gene Ontology Annotation (GOA) Database

The GO annotation program aims to provide high-quality Gene Ontology (GO) annotations to proteins in the UniProt Knowledgebase (UniProtKB), RNA molecules from RNACentral and protein complexes from the Complex Portal. To search and view Gene Ontology terms and annotations, please use our QuickGO browser.

GOA files contain a mixture of manual annotation supplied by members of the Gene Onotology Consortium and computationally assigned GO terms describing gene products. Annotation type is clearly indicated by associated evidence codes and there are links to the source data.

Release cycle

All GOA files are released approximately every four weeks and can be accessed from our Downloads page. We aim to coordinate our release with releases of UniProtKB.

Latest statistics

UniProt >	Human >
Mouse >	Rat >
Arabidopsis >	Zebrafish >
Chicken >	Cow >
Dicty >	Dog >
Pig >	Fly >
Worm >	Yeast >
Proteomes >	

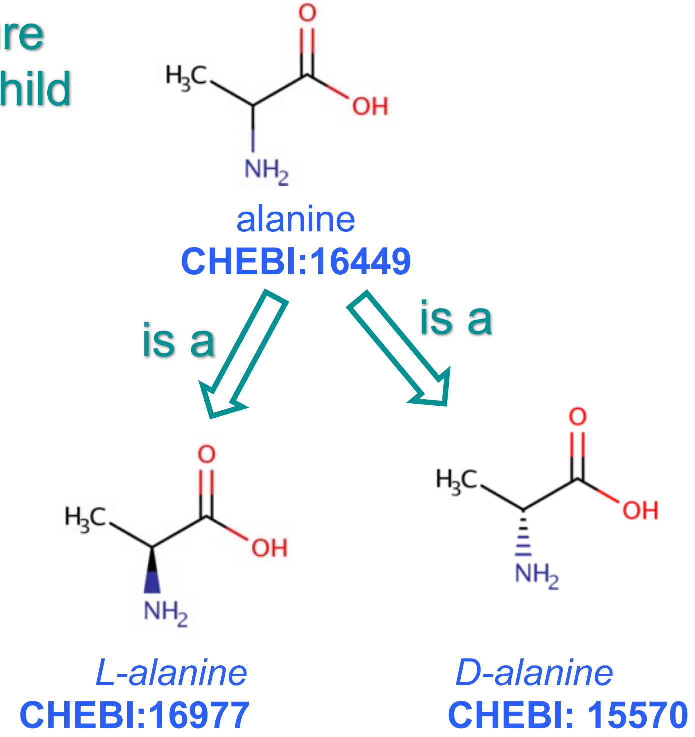
The GOA relationships between UniProt and GO are most often many-to-many

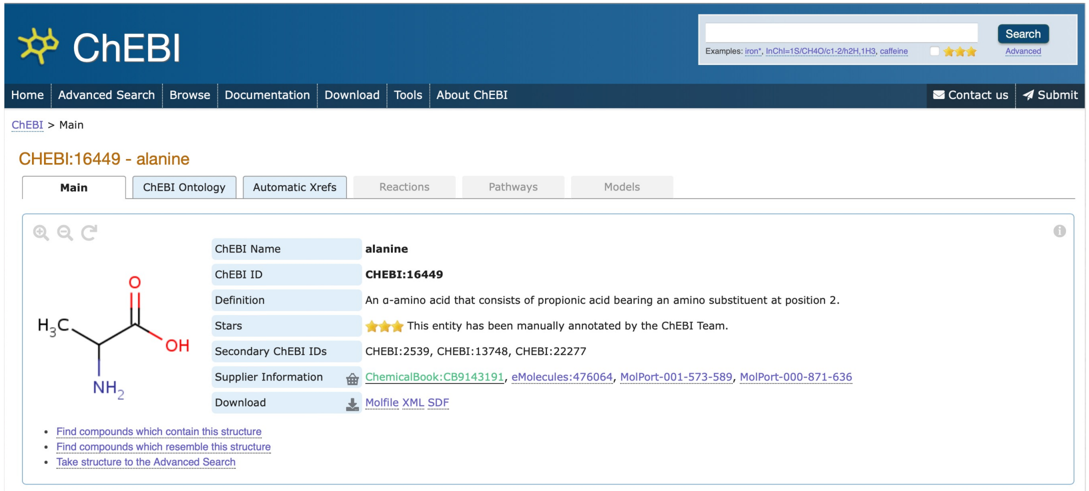
They are readily available for model organisms

They can be computed for non-model organisms, using InterProScan for example. The resulting annotations are often too general to lead to interesting enrichment results.

ChEBI: Chemical Entities of Biological Interest

The parent structure encompasses the child structure





ChEBI

Examples: iron*, InChI=1S/CH4O/c1-2/h2H,1H3, caffeine □ ★★★

Search

Advanced

Home | Advanced Search | Browse | Documentation | Download | Tools | About ChEBI | Contact us | Submit

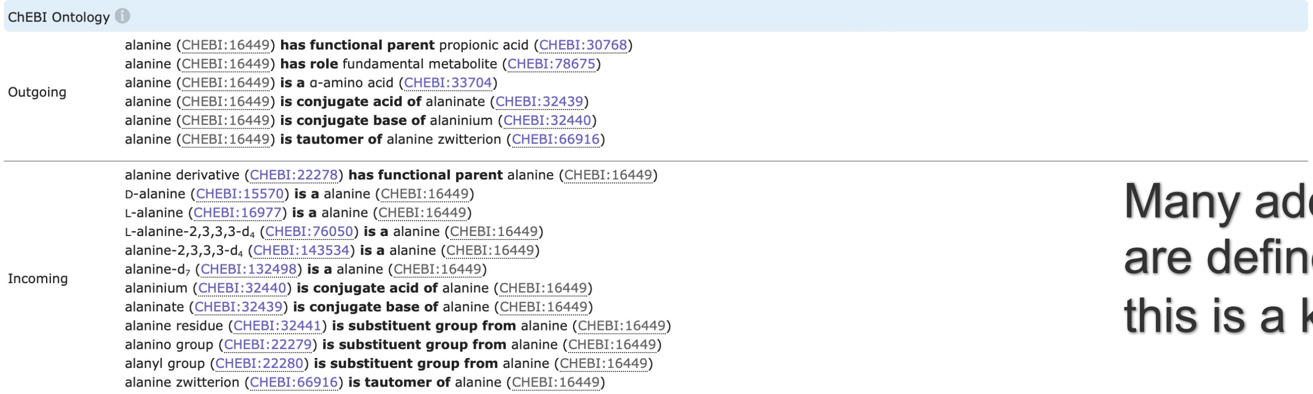
ChEBI > Main

CHEBI:16449 - alanine

Main | ChEBI Ontology | Automatic Xrefs | Reactions | Pathways | Models

ChEBI Name	**alanine**
ChEBI ID	**CHEBI:16449**
Definition	An α-amino acid that consists of propionic acid bearing an amino substituent at position 2.
Stars	★★★ This entity has been manually annotated by the ChEBI Team.
Secondary ChEBI IDs	CHEBI:2539, CHEBI:13748, CHEBI:22277
Supplier Information	ChemicalBook:CB9143191, eMolecules:476064, MolPort-001-573-589, MolPort-000-871-636
Download	Molfile XML SDF

- Find compounds which contain this structure
- Find compounds which resemble this structure
- Take structure to the Advanced Search



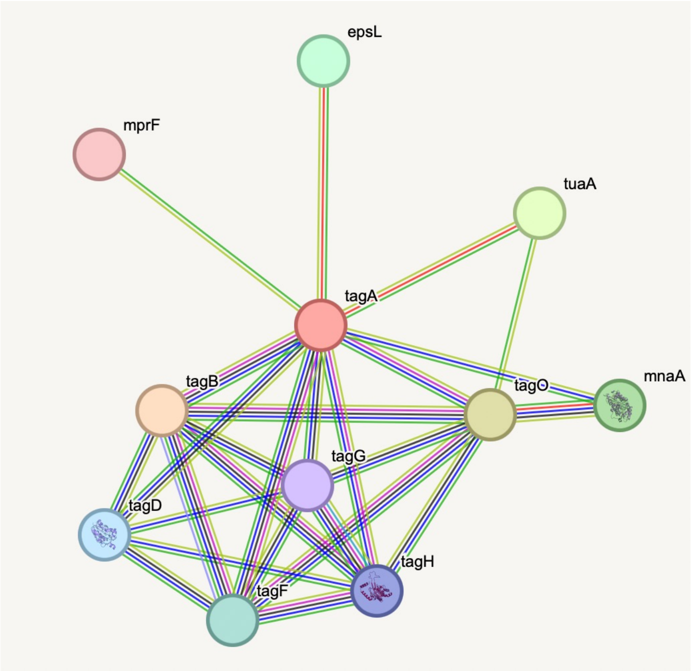
ChEBI Ontology

Outgoing	alanine (CHEBI:16449) **has functional parent** propionic acid (CHEBI:30768) alanine (CHEBI:16449) **has role** fundamental metabolite (CHEBI:78675) alanine (CHEBI:16449) **is a** α-amino acid (CHEBI:33704) alanine (CHEBI:16449) **is conjugate acid of** alaninate (CHEBI:32439) alanine (CHEBI:16449) **is conjugate base of** alaninium (CHEBI:32440) alanine (CHEBI:16449) **is tautomer of** alanine zwitterion (CHEBI:66916)
Incoming	alanine derivative (CHEBI:22278) **has functional parent** alanine (CHEBI:16449) D-alanine (CHEBI:15570) **is a** alanine (CHEBI:16449) L-alanine (CHEBI:16977) **is a** alanine (CHEBI:16449) L-alanine-2,3,3,3-d_4 (CHEBI:76050) **is a** alanine (CHEBI:16449) alanine-2,3,3,3-d_4 (CHEBI:143534) **is a** alanine (CHEBI:16449) alanine-d_7 (CHEBI:132498) **is a** alanine (CHEBI:16449) alaninium (CHEBI:32440) **is conjugate acid of** alanine (CHEBI:16449) alaninate (CHEBI:32439) **is conjugate base of** alanine (CHEBI:16449) alanine residue (CHEBI:32441) **is substituent group from** alanine (CHEBI:16449) alanino group (CHEBI:22279) **is substituent group from** alanine (CHEBI:16449) alanyl group (CHEBI:22280) **is substituent group from** alanine (CHEBI:16449) alanine zwitterion (CHEBI:66916) **is tautomer of** alanine (CHEBI:16449)

Many additional relationships are defined among entries: this is a knowledge graph

Database structure

STRINGdb is primarily an undirected graph with different types of relationships (different supporting evidences) among entries



Notations for external identifiers

The most commonly used compact notation nowadays is

prefix:accession-number

In the RDF world using Turtle syntax, given

```
PREFIX up: <http://purl.uniprot.org/uniprot/>
```

the identifier in short form

```
up:P29358
```

refers exactly to the same entity as the identifier in long form

```
<http://purl.uniprot.org/uniprot/P29358>
```

In RDF, the prefix definition is local, not public. Hence the stable public identifier is the long form.

Difficulties with prefix nomenclatures

Utilisation of prefixes is well codified in the RDF world. Less elsewhere!

`identifiers.org` is attempting to promote universal public prefixes. Unfortunately they have created new long forms, ignoring widely-used previous ones. This has introduced unnecessary communication difficulties between the Systems Biology and Bioinformatics communities.

Nota Bene: some identifiers were originally defined with a ":" as part of the accession numbers. In practice:

`CHEBI:16977`, `chebi:CHEBI:16977` and `chebi:CHEBI_16977` are very likely to refer to the same entry in ChEBI

Cross reference semantics

What does a cross reference means?

- Different identifiers for exactly the same entity
- Different identifiers for closely-related object, *e.g. in* two database of protein, one with the focus on protein structures, the other one on protein sequences
- Different identifiers for distinct but related object
  - gene ⇔ protein

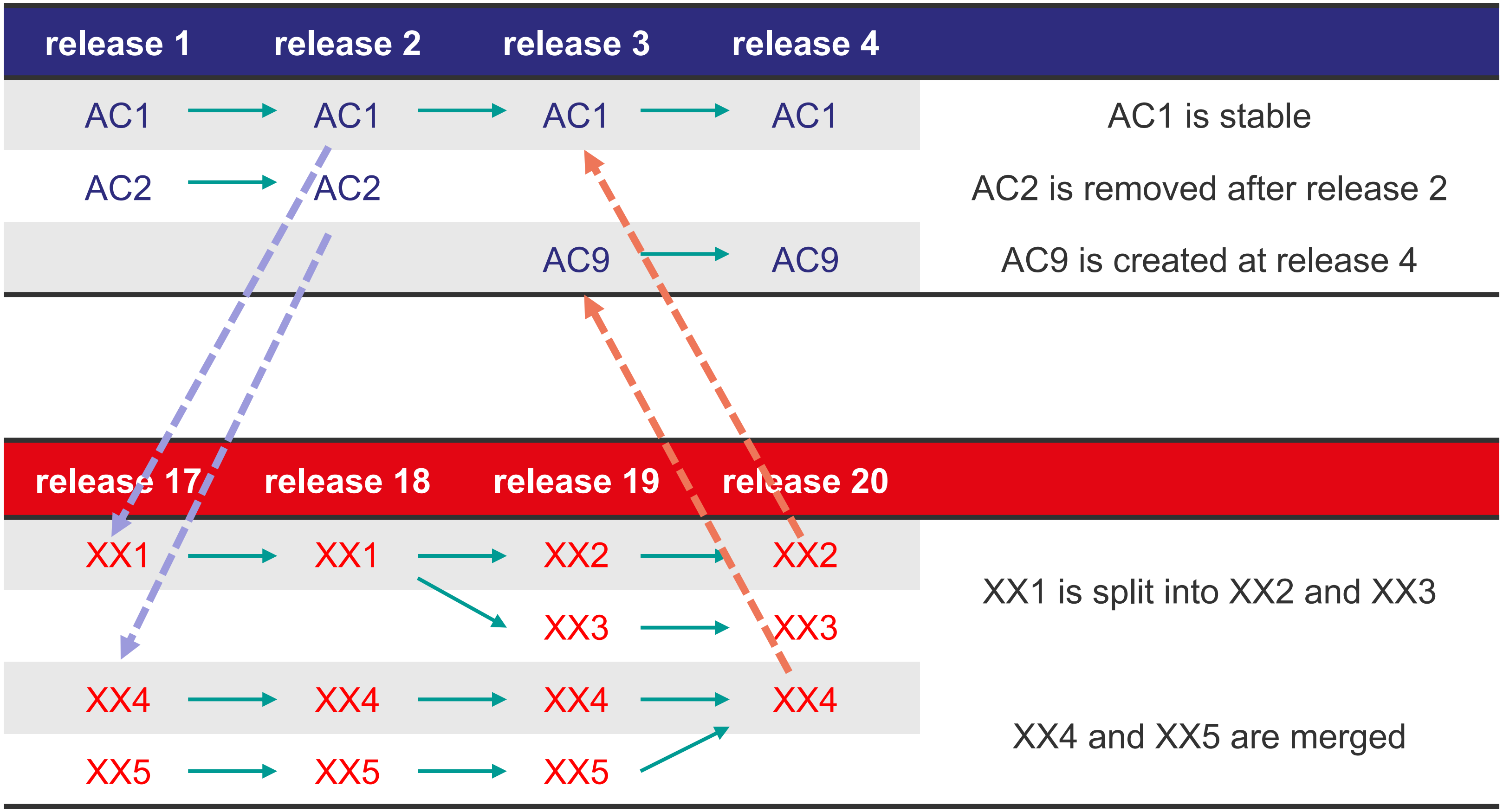
Cross references

A cross reference links an entry in a database to another entry in another database, *i.e.* using an external accession number

```
DR   EMBL; BC120112; AAI20113.1; -; mRNA.
DR   EMBL; AF043736; AAC02090.1; -; mRNA.
DR   PIR; S13467; S13467.
DR   RefSeq; NP_777219.2; NM_174794.2.
DR   AlphaFoldDB; P68250; -.
DR   SMR; P68250; -.
DR   STRING; 9913.ENSBTAP00000022411; -.
DR   iPTMnet; P68250; -.
DR   PaxDb; P68250; -.
DR   PeptideAtlas; P68250; -.
DR   GeneID; 286863; -.
DR   KEGG; bta:286863; -.
DR   GO; GO:0005737; C:cytoplasm; ISS:AgBase.
DR   GO; GO:0042470; C:melanosome; IEA:UniProtKB-SubCell.
DR   GO; GO:0048471; C:perinuclear region of cytoplasm; ISS:AgBase.
DR   GO; GO:0019904; F:protein domain specific binding; ISS:AgBase.
DR   InterPro; IPR000308; 14-3-3.
DR   InterPro; IPR023409; 14-3-3_CS.
DR   InterPro; IPR036815; 14-3-3_dom_sf.
DR   InterPro; IPR023410; 14-3-3_domain.
DR   PANTHER; PTHR18860; 14-3-3 PROTEIN; 1.
DR   PANTHER; PTHR18860:SF28; 14-3-3 PROTEIN BETA/ALPHA; 1.
```

Cross references across releases

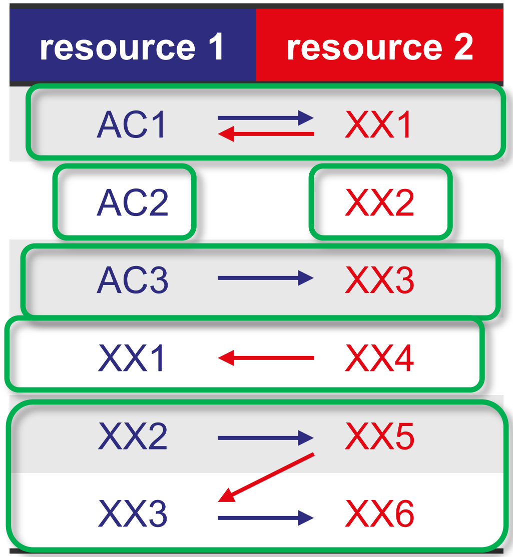
For practical reason, cross-references usually refer to entries in previous release of the external databases !



release 1	release 2	release 3	release 4	
AC1	AC1	AC1	AC1	AC1 is stable
AC2	AC2			AC2 is removed after release 2
		AC9	AC9	AC9 is created at release 4

release 17	release 18	release 19	release 20	
XX1	XX1	XX2	XX2	XX1 is split into XX2 and XX3
		XX3	XX3	
XX4	XX4	XX4	XX4	XX4 and XX5 are merged
XX5	XX5	XX5		

Cross references



resource 1		resource 2
AC1	⇄	XX1
AC2		XX2
AC3	→	XX3
XX1	←	XX4
XX2	→	XX5
XX3	→	XX6

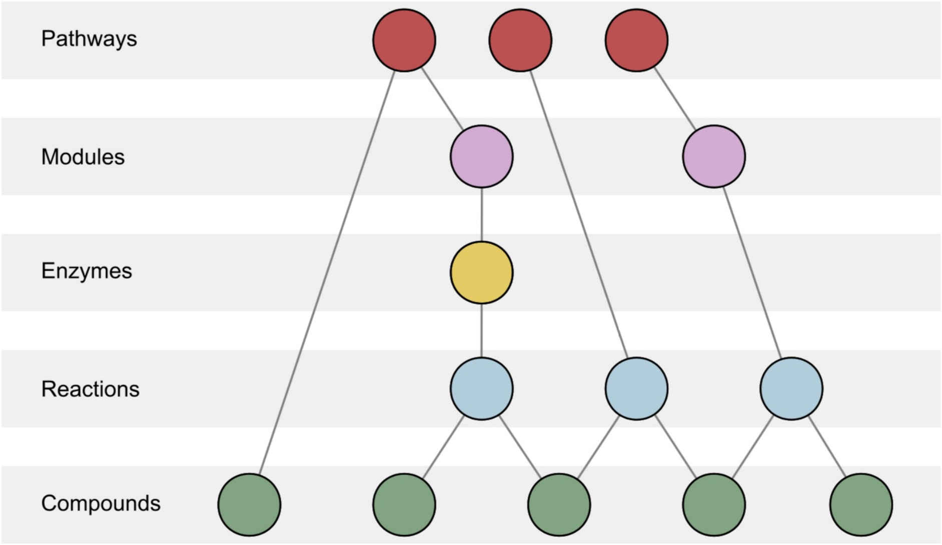
Computation of connected component helps to understand structure of cross references (available from `igraph`)

Database structure

KEGG is a DAG with a fixed depth.

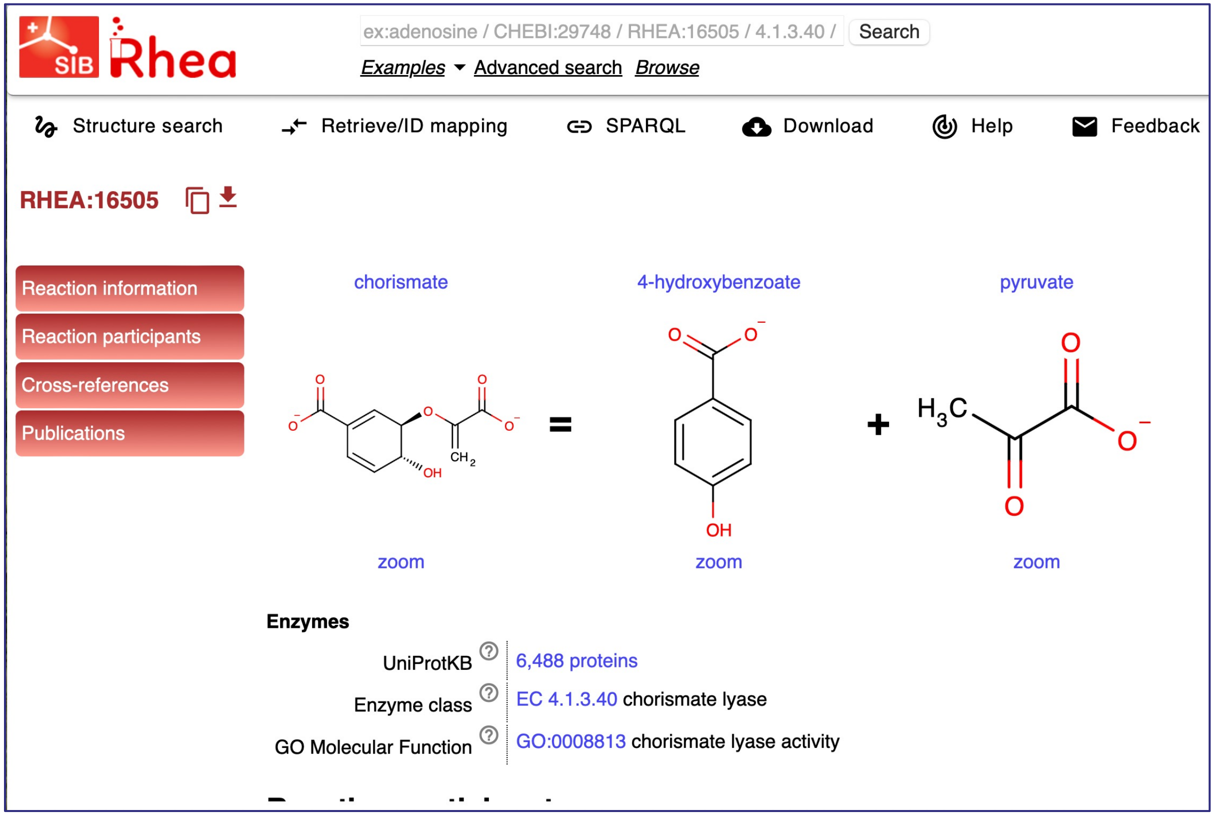
It can also be viewed as a multipartite graph. Individual entries are assigned a type (compound, reaction ...) and relations are only considered between entries of different types.

The complete description of a biochemical pathways down to the metabolite level is the induced sub-graph starting from the pathway identifier.



RHEA

A database of biochemical reaction meant to annotate enzymes in UniProt



RHEA:16505

chorismate = 4-hydroxybenzoate + pyruvate

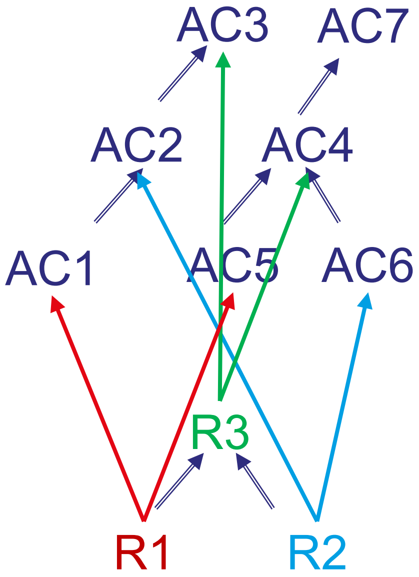
Enzymes

UniProtKB	6,488 proteins
Enzyme class	EC 4.1.3.40 chorismate lyase
GO Molecular Function	GO:0008813 chorismate lyase activity

RHEA

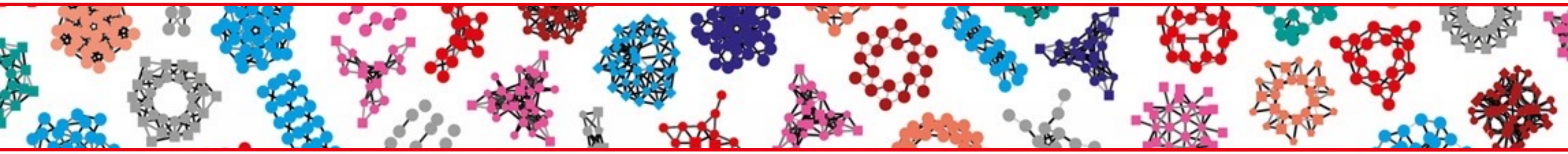
RHEA is an ontology (a DAG) of biochemical reactions with reactants taken from the ChEBI ontology (another DAG)

R1: AC1 ⇔ AC5
R2: AC2 ⇔ AC6
R3: AC3 ⇔ AC4



Reasoning with ontologies:

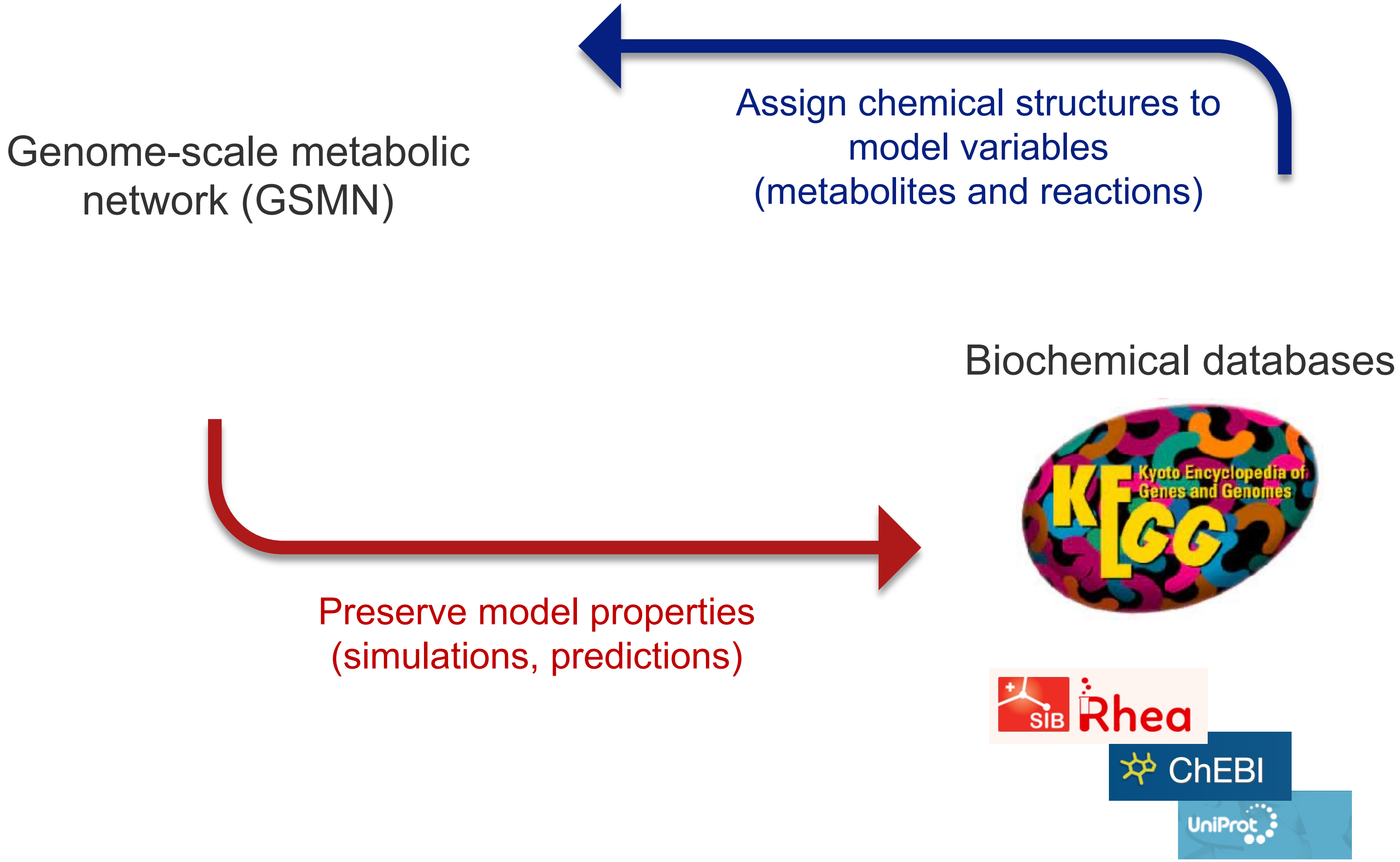
1. AC1 ⇔ AC6 is implied by R3
2. reaction R1 is a child of R2



Adding molecular structures to stoichiometric models:

...where Systems Biology and Chemoinformatics meet

An old problem



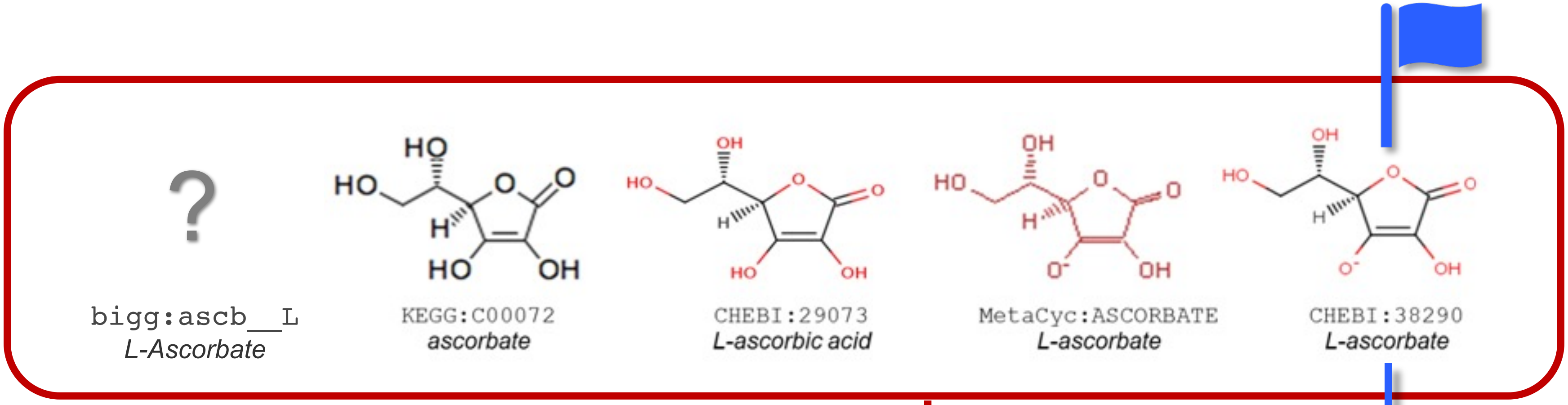
Genome-scale metabolic network (GSMN)

Assign chemical structures to model variables (metabolites and reactions)

Biochemical databases

Preserve model properties (simulations, predictions)

Merging chemicals / selecting representative



?

bigg:ascb__L
L-Ascorbate

KEGG:C00072
ascorbate

CHEBI:29073
L-ascorbic acid

MetaCyc:ASCORBATE
L-ascorbate

CHEBI:38290
L-ascorbate

Evidences to group metabolites (by decreasing importance):

1. Chemical records
2. Reaction contexts
3. Cross-references (with care)
4. Names

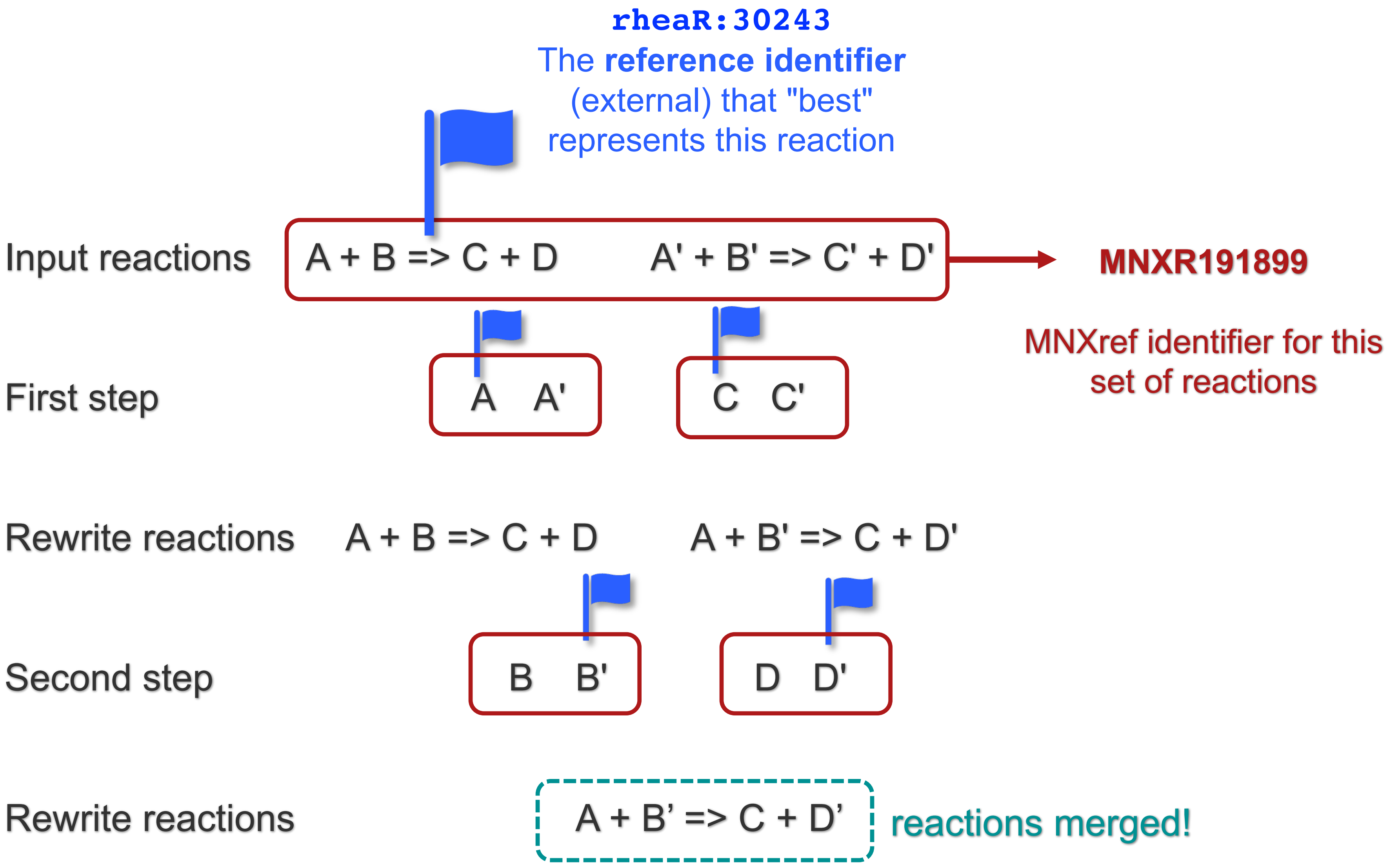
MNXM727871

The MNXref identifier for this metabolite, *i.e.* an identifier for the set of molecules that are grouped together

chebi:38290

The **reference** (external) identifier that "best" represents this metabolite

Merging chemicals => merging reactions



rheaR:30243

The **reference identifier** (external) that "best" represents this reaction

Input reactions A + B => C + D A' + B' => C' + D'

MNXR191899

MNXref identifier for this set of reactions

First step A A' C C'

Rewrite reactions A + B => C + D A + B' => C + D'

Second step B B' D D'

Rewrite reactions A + B’ => C + D’

reactions merged!

MNXref reconciliation in number (release 4.2)

metabolites

	all	in reac	in mnet*
CHEBI	116222	21196	5637
bigg	9130	9053	6541
envipath	12306	1580	591
hmdb	195008	9369	4500
keggC	18673	9899	2757
keggD	11147	650	250
keggE	864	0	0
keggG	11042	406	115
lipidmaps	43085	2832	929
metacyc	20296	16199	2505
reactome	5526	2031	1638
sabiork	8944	8899	1443
seed	33995	21634	3789
slm	777657	1831	524
MNXref	**1043605**	**41584**	**9359**
ratio	**2.15**	**3.62**	**5.16**

*approx. 150 public GEMs from different labs and different organisms

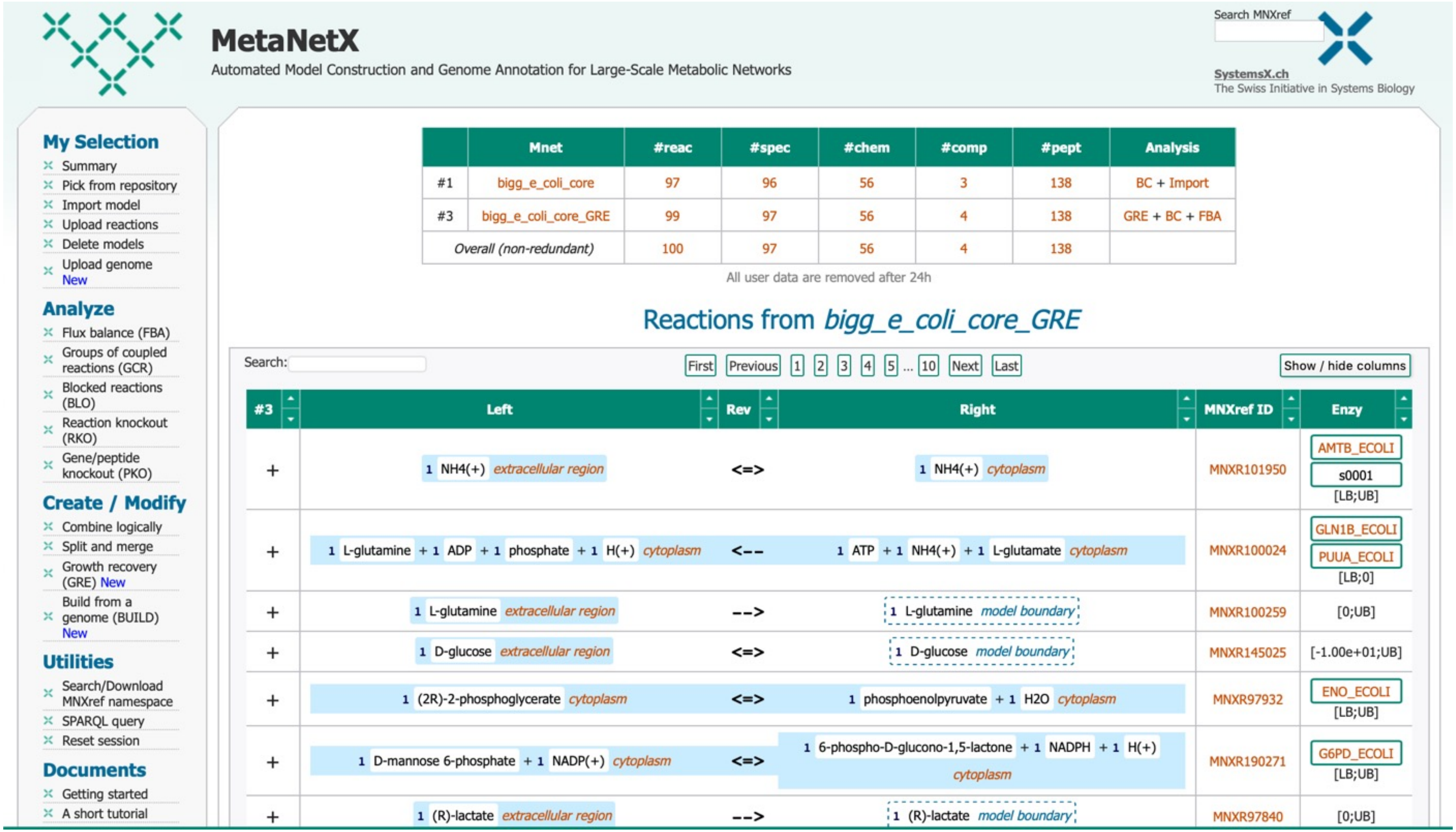
MetaNetX
Automated Model Construction
and Genome Annotation for
Large-Scale Metabolic Networks

reactions

	all	in mnet
bigg	28167	40653
kegg	11160	2879
metacyc	17198	3262
rhea	12510	3101
sabiork	8118	1818
seed	43855	15958
MNXref	**36944**	**13317**
ratio	**4.63**	**6.38**

The full dataset is distributed in TAB-delimited and RDF/Turtle formats under CC-BY license

Navigating MNXref at `www.metanetx.org`



MetaNetX

Automated Model Construction and Genome Annotation for Large-Scale Metabolic Networks

Search MNXref

SystemsX.ch

The Swiss Initiative in Systems Biology

My Selection

- Summary
- Pick from repository
- Import model
- Upload reactions
- Delete models
- Upload genome New

Analyze

- Flux balance (FBA)
- Groups of coupled reactions (GCR)
- Blocked reactions (BLO)
- Reaction knockout (RKO)
- Gene/peptide knockout (PKO)

Create / Modify

- Combine logically
- Split and merge
- Growth recovery (GRE) New
- Build from a genome (BUILD) New

Utilities

- Search/Download MNXref namespace
- SPARQL query
- Reset session

Documents

- Getting started
- A short tutorial

	Mnet	#reac	#spec	#chem	#comp	#pept	Analysis
#1	bigg_e_coli_core	97	96	56	3	138	BC + Import
#3	bigg_e_coli_core_GRE	99	97	56	4	138	GRE + BC + FBA
	Overall (non-redundant)	100	97	56	4	138	

All user data are removed after 24h

Reactions from *bigg_e_coli_core_GRE*

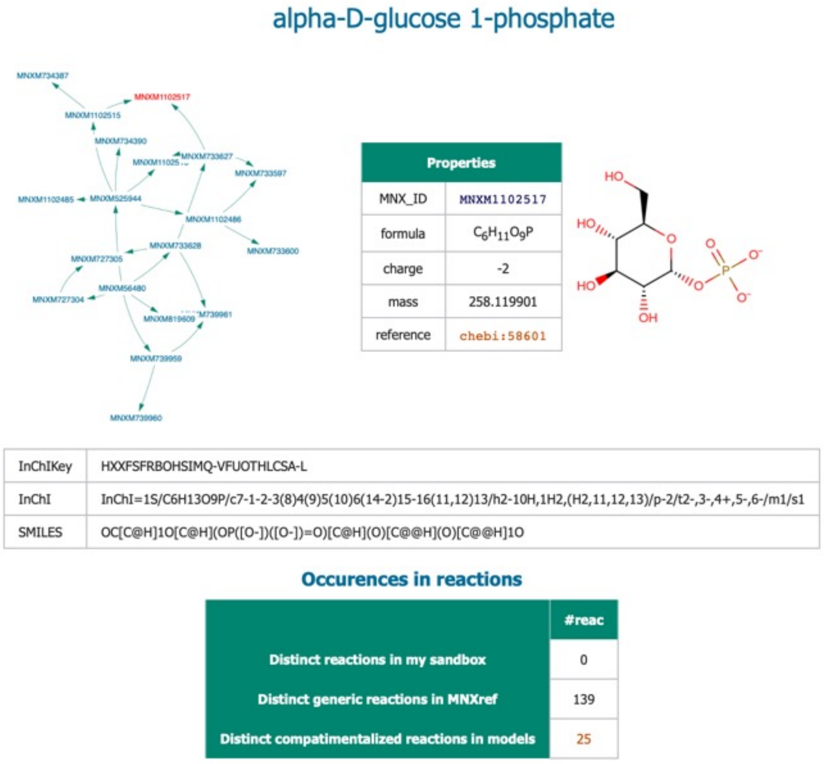
Search:

First Previous 1 2 3 4 5 ... 10 Next Last

Show / hide columns

#3	Left	Rev	Right	MNXref ID	Enzy
+	1 NH4(+) *extracellular region*	<=>	1 NH4(+) *cytoplasm*	MNXR101950	AMTB_ECOLI s0001 [LB;UB]
+	1 L-glutamine + 1 ADP + 1 phosphate + 1 H(+) *cytoplasm*	<--	1 ATP + 1 NH4(+) + 1 L-glutamate *cytoplasm*	MNXR100024	GLN1B_ECOLI PUUA_ECOLI [LB;0]
+	1 L-glutamine *extracellular region*	-->	1 L-glutamine *model boundary*	MNXR100259	[0;UB]
+	1 D-glucose *extracellular region*	<=>	1 D-glucose *model boundary*	MNXR145025	[-1.00e+01;UB]
+	1 (2R)-2-phosphoglycerate *cytoplasm*	<=>	1 phosphoenolpyruvate + 1 H2O *cytoplasm*	MNXR97932	ENO_ECOLI [LB;UB]
+	1 D-mannose 6-phosphate + 1 NADP(+) *cytoplasm*	<=>	1 6-phospho-D-glucono-1,5-lactone + 1 NADPH + 1 H(+) *cytoplasm*	MNXR190271	G6PD_ECOLI [LB;UB]
+	1 (R)-lactate *extracellular region*	-->	1 (R)-lactate *model boundary*	MNXR97840	[0;UB]

Navigating MNXref at **www.metanetx.org**



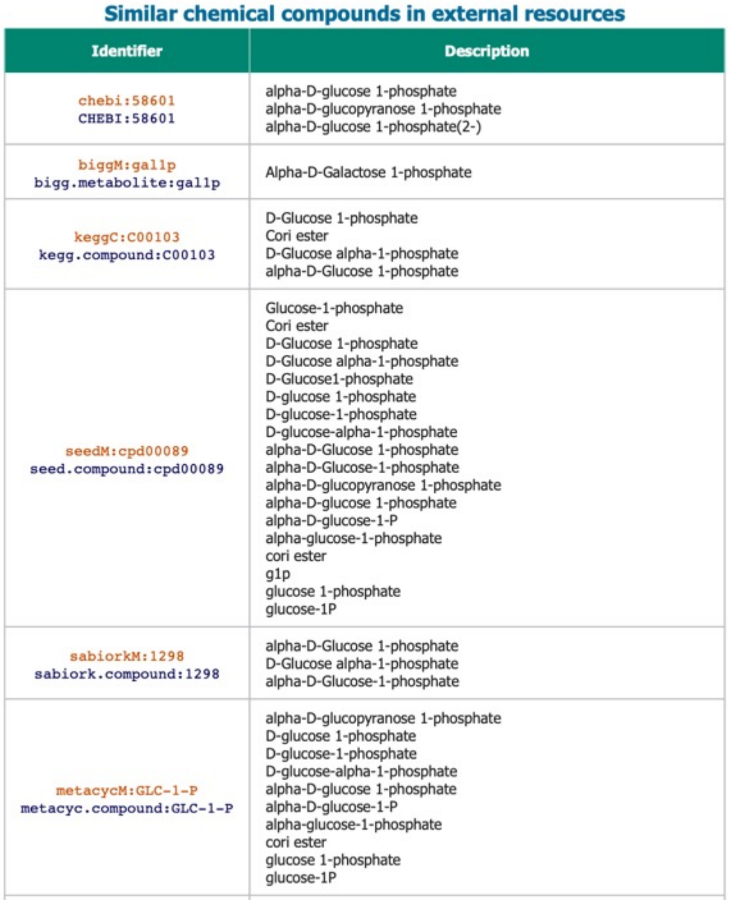
alpha-D-glucose 1-phosphate

Properties	
MNX_ID	MNXM1102517
formula	$C_6H_{11}O_9P$
charge	-2
mass	258.119901
reference	chebi:58601

InChIKey	HXXFSFRBOHSIMQ-VFUOTHLCSA-L
InChI	InChI=1S/C6H13O9P/c7-1-2-3(8)4(9)5(10)6(14-2)15-16(11,12)13/h2-10H,1H2,(H2,11,12,13)/p-2/t2-,3-,4+,5-,6-/m1/s1
SMILES	OC[C@H]1O[C@H](OP([O-])([O-])=O)[C@H](O)[C@@H](O)[C@@H]1O

Occurences in reactions

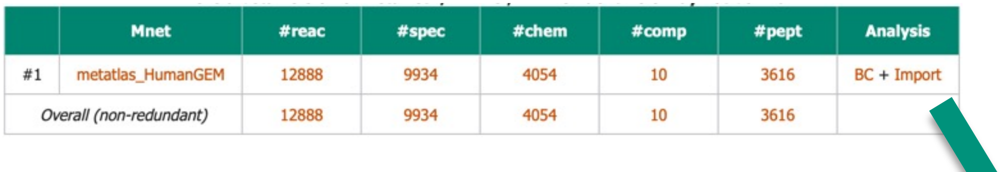
	#reac
Distinct reactions in my sandbox	0
Distinct generic reactions in MNXref	139
Distinct compatimentalized reactions in models	25



Similar chemical compounds in external resources

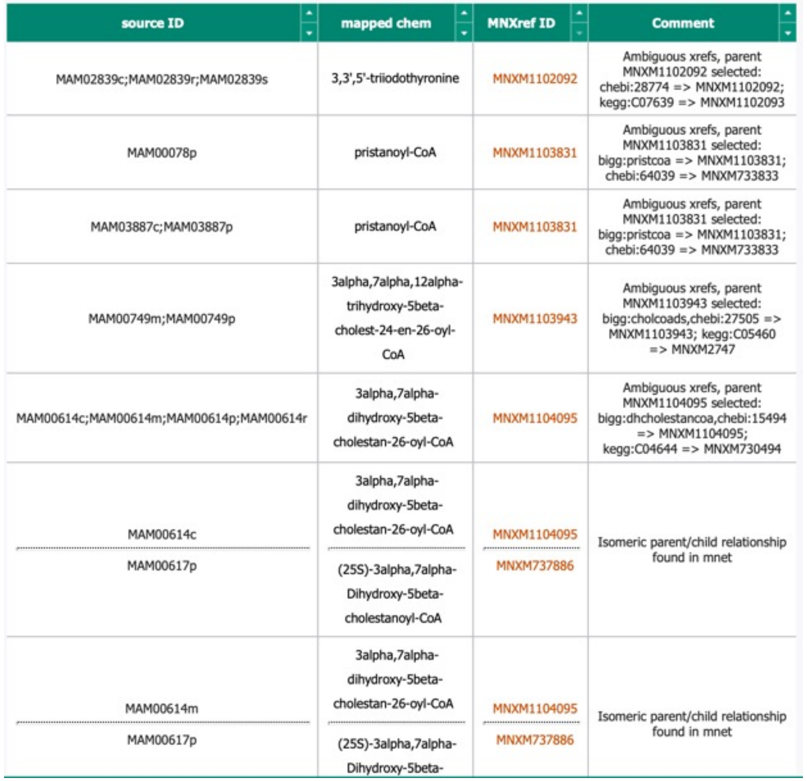
Identifier	Description
chebi:58601 CHEBI:58601	alpha-D-glucose 1-phosphate alpha-D-glucopyranose 1-phosphate alpha-D-glucose 1-phosphate(2-)
biggM:gal1p bigg.metabolite:gal1p	Alpha-D-Galactose 1-phosphate
keggC:C00103 kegg.compound:C00103	D-Glucose 1-phosphate Cori ester D-Glucose alpha-1-phosphate alpha-D-Glucose 1-phosphate
seedM:cpd00089 seed.compound:cpd00089	Glucose-1-phosphate Cori ester D-Glucose 1-phosphate D-Glucose alpha-1-phosphate D-Glucose1-phosphate D-glucose 1-phosphate D-glucose-1-phosphate D-glucose-alpha-1-phosphate alpha-D-Glucose 1-phosphate alpha-D-Glucose-1-phosphate alpha-D-glucopyranose 1-phosphate alpha-D-glucose 1-phosphate alpha-D-glucose-1-P alpha-glucose-1-phosphate cori ester g1p glucose 1-phosphate glucose-1P
sabiorkM:1298 sabiork.compound:1298	alpha-D-Glucose 1-phosphate D-Glucose alpha-1-phosphate alpha-D-Glucose-1-phosphate
metacycM:GLC-1-P metacyc.compound:GLC-1-P	alpha-D-glucopyranose 1-phosphate D-glucose 1-phosphate D-glucose-1-phosphate D-glucose-alpha-1-phosphate alpha-D-glucose 1-phosphate alpha-D-glucose-1-P alpha-glucose-1-phosphate cori ester glucose 1-phosphate glucose-1P

Diagnose metabolic networks



	Mnet	#reac	#spec	#chem	#comp	#pept	Analysis
#1	metatlas_HumanGEM	12888	9934	4054	10	3616	BC + Import
Overall (non-redundant)		12888	9934	4054	10	3616	

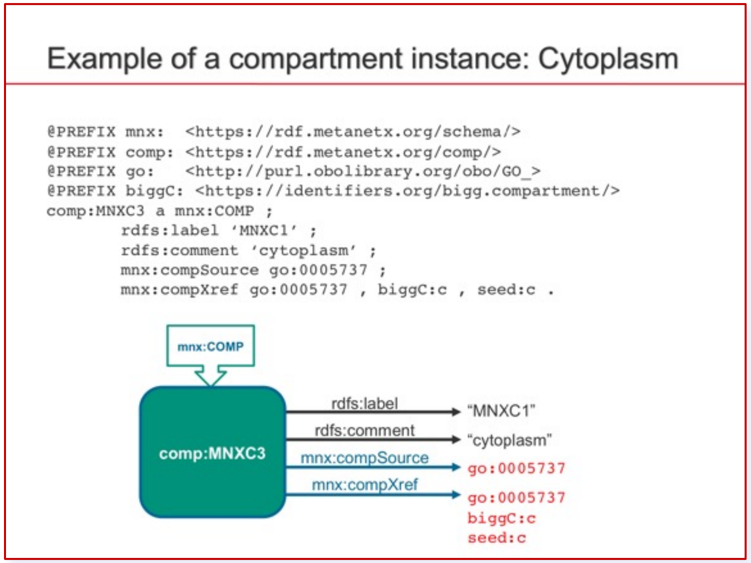
- ambiguous and conflicting mapping to MNXref
- duplicated reactions
- metabolites with isomeric parent/child relationships



source ID	mapped chem	MNXref ID	Comment
MAM02839c;MAM02839r;MAM02839s	3,3',5'-triiodothyronine	MNXM1102092	Ambiguous xrefs, parent MNXM1102092 selected: chebi:28774 => MNXM1102092; kegg:C07639 => MNXM1102093
MAM00078p	pristanoyl-CoA	MNXM1103831	Ambiguous xrefs, parent MNXM1103831 selected: bigg:pristcoa => MNXM1103831; chebi:64039 => MNXM733833
MAM03887c;MAM03887p	pristanoyl-CoA	MNXM1103831	Ambiguous xrefs, parent MNXM1103831 selected: bigg:pristcoa => MNXM1103831; chebi:64039 => MNXM733833
MAM00749m;MAM00749p	3alpha,7alpha,12alpha-trihydroxy-5beta-cholest-24-en-26-oyl-CoA	MNXM1103943	Ambiguous xrefs, parent MNXM1103943 selected: bigg:cholcoads,chebi:27505 => MNXM1103943; kegg:C05460 => MNXM2747
MAM00614c;MAM00614m;MAM00614p;MAM00614r	3alpha,7alpha-dihydroxy-5beta-cholestan-26-oyl-CoA	MNXM1104095	Ambiguous xrefs, parent MNXM1104095 selected: bigg:dhcholestancoa,chebi:15494 => MNXM1104095; kegg:C04644 => MNXM730494
MAM00614c MAM00617p	3alpha,7alpha-dihydroxy-5beta-cholestan-26-oyl-CoA (25S)-3alpha,7alpha-Dihydroxy-5beta-cholestanoyl-CoA	MNXM1104095 MNXM737886	Isomeric parent/child relationship found in mnet
MAM00614m MAM00617p	3alpha,7alpha-dihydroxy-5beta-cholestan-26-oyl-CoA (25S)-3alpha,7alpha-Dihydroxy-5beta-	MNXM1104095 MNXM737886	Isomeric parent/child relationship found in mnet

RDF/Turtle distribution and SPARQL endpoint

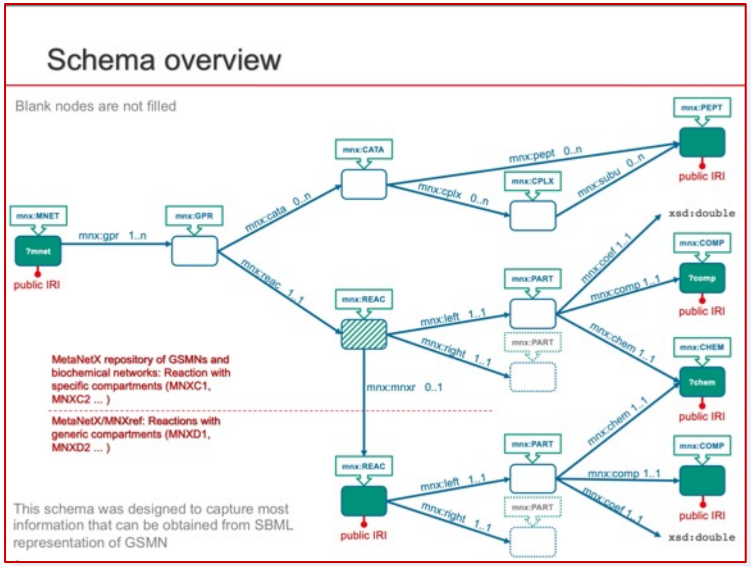
https://rdf.metanetx.org



Example of a compartment instance: Cytoplasm

```
@PREFIX mnx:   <https://rdf.metanetx.org/schema/>
@PREFIX comp: <https://rdf.metanetx.org/comp/>
@PREFIX go:    <http://purl.obolibrary.org/obo/GO_>
@PREFIX biggC: <https://identifiers.org/bigg.compartment/>
comp:MNXC3 a mnx:COMP ;
        rdfs:label ‘MNXC1’ ;
        rdfs:comment ‘cytoplasm’ ;
        mnx:compSource go:0005737 ;
        mnx:compXref go:0005737 , biggC:c , seed:c .
```

mnx:COMP
comp:MNXC3
rdfs:label
"MNXC1"
rdfs:comment
"cytoplasm"
mnx:compSource
go:0005737
mnx:compXref
go:0005737
biggC:c
seed:c



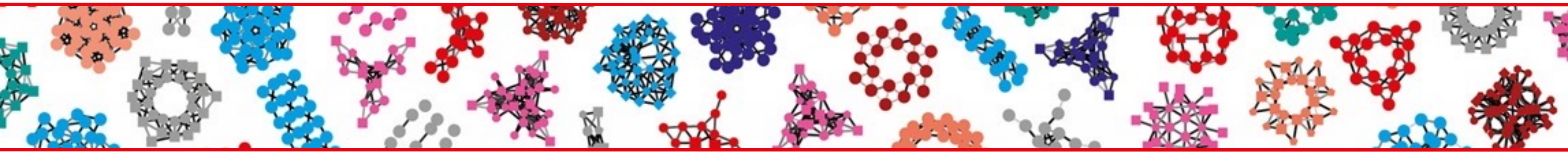
Schema overview

Blank nodes are not filled

MetaNetX repository of GSMNs and biochemical networks: Reaction with specific compartments (MNXC1, MNXC2 ...)

MetaNetX/MNXref: Reactions with generic compartments (MNXD1, MNXD2 ...)

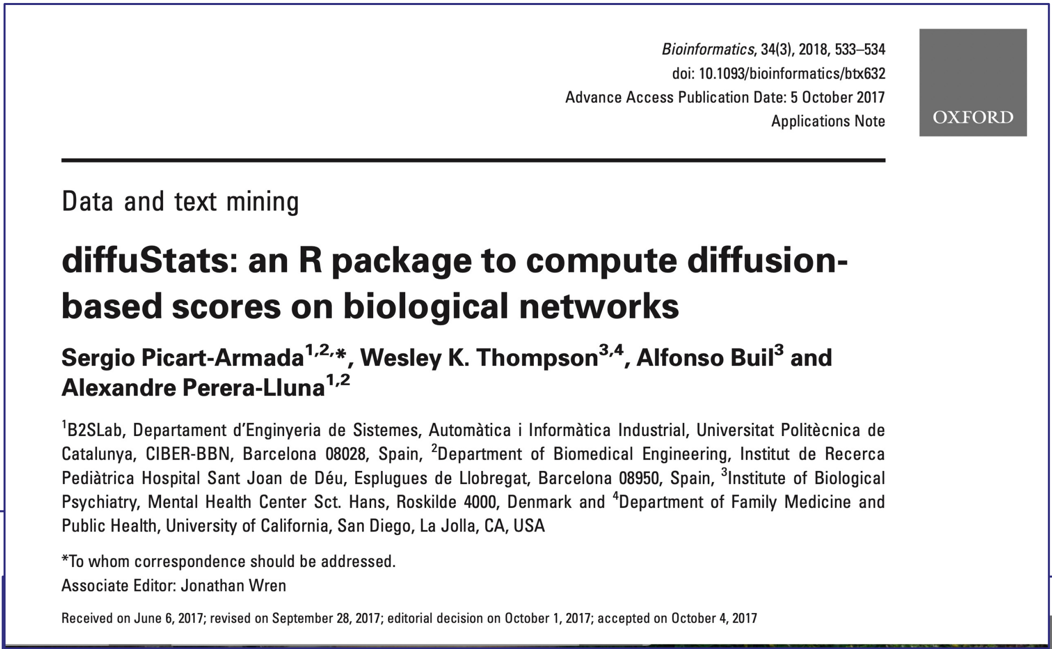
This schema was designed to capture most information that can be obtained from SBML representation of GSMN



Graph diffusion

R packages from Bioconductor:

- FELLA
- diffuStats



Bioinformatics, 34(3), 2018, 533–534
doi: 10.1093/bioinformatics/btx632
Advance Access Publication Date: 5 October 2017
Applications Note

Data and text mining

diffuStats: an R package to compute diffusion-based scores on biological networks

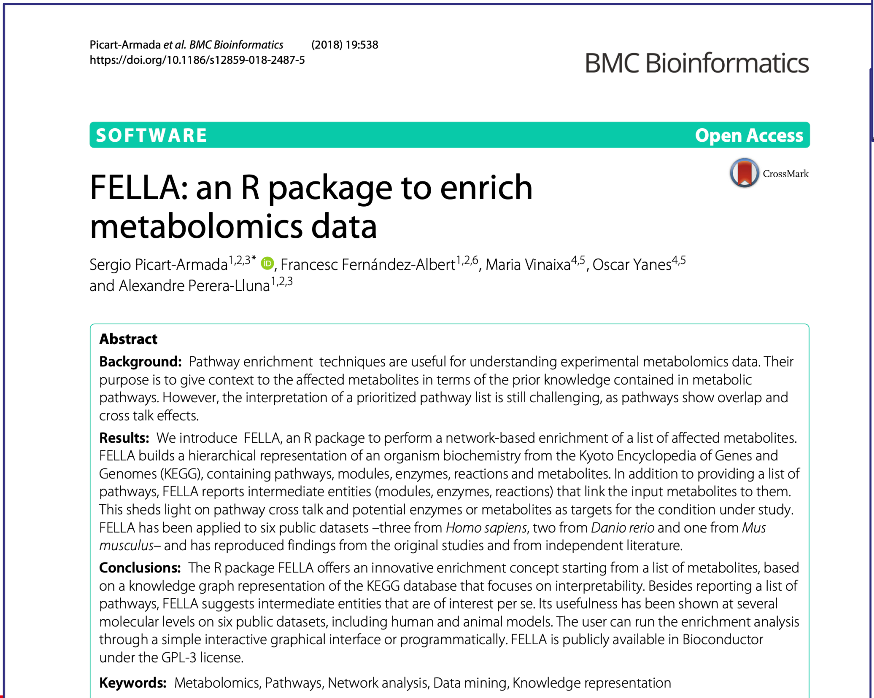
Sergio Picart-Armada[1,2,*], Wesley K. Thompson[3,4], Alfonso Buil[3] and Alexandre Perera-Lluna[1,2]

[1]B2SLab, Departament d'Enginyeria de Sistemes, Automàtica i Informàtica Industrial, Universitat Politècnica de Catalunya, CIBER-BBN, Barcelona 08028, Spain, [2]Department of Biomedical Engineering, Institut de Recerca Pediàtrica Hospital Sant Joan de Déu, Esplugues de Llobregat, Barcelona 08950, Spain, [3]Institute of Biological Psychiatry, Mental Health Center Sct. Hans, Roskilde 4000, Denmark and [4]Department of Family Medicine and Public Health, University of California, San Diego, La Jolla, CA, USA

*To whom correspondence should be addressed.

Associate Editor: Jonathan Wren

Received on June 6, 2017; revised on September 28, 2017; editorial decision on October 1, 2017; accepted on October 4, 2017



Picart-Armada *et al. BMC Bioinformatics* (2018) 19:538
https://doi.org/10.1186/s12859-018-2487-5

BMC Bioinformatics

SOFTWARE Open Access

FELLA: an R package to enrich metabolomics data

Sergio Picart-Armada[1,2,3*], Francesc Fernández-Albert[1,2,6], Maria Vinaixa[4,5], Oscar Yanes[4,5] and Alexandre Perera-Lluna[1,2,3]

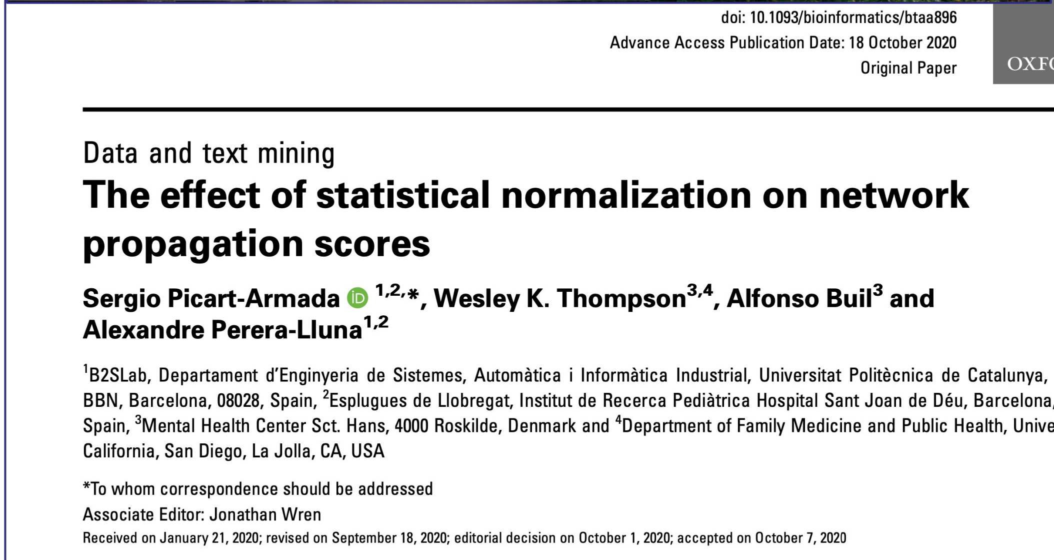
Abstract

Background: Pathway enrichment techniques are useful for understanding experimental metabolomics data. Their purpose is to give context to the affected metabolites in terms of the prior knowledge contained in metabolic pathways. However, the interpretation of a prioritized pathway list is still challenging, as pathways show overlap and cross talk effects.

Results: We introduce FELLA, an R package to perform a network-based enrichment of a list of affected metabolites. FELLA builds a hierarchical representation of an organism biochemistry from the Kyoto Encyclopedia of Genes and Genomes (KEGG), containing pathways, modules, enzymes, reactions and metabolites. In addition to providing a list of pathways, FELLA reports intermediate entities (modules, enzymes, reactions) that link the input metabolites to them. This sheds light on pathway cross talk and potential enzymes or metabolites as targets for the condition under study. FELLA has been applied to six public datasets –three from *Homo sapiens*, two from *Danio rerio* and one from *Mus musculus*– and has reproduced findings from the original studies and from independent literature.

Conclusions: The R package FELLA offers an innovative enrichment concept starting from a list of metabolites, based on a knowledge graph representation of the KEGG database that focuses on interpretability. Besides reporting a list of pathways, FELLA suggests intermediate entities that are of interest per se. Its usefulness has been shown at several molecular levels on six public datasets, including human and animal models. The user can run the enrichment analysis through a simple interactive graphical interface or programmatically. FELLA is publicly available in Bioconductor under the GPL-3 license.

Keywords: Metabolomics, Pathways, Network analysis, Data mining, Knowledge representation



doi: 10.1093/bioinformatics/btaa896
Advance Access Publication Date: 18 October 2020
Original Paper

Data and text mining

The effect of statistical normalization on network propagation scores

Sergio Picart-Armada [1,2,*], Wesley K. Thompson[3,4], Alfonso Buil[3] and Alexandre Perera-Lluna[1,2]

[1]B2SLab, Departament d'Enginyeria de Sistemes, Automàtica i Informàtica Industrial, Universitat Politècnica de Catalunya, BBN, Barcelona, 08028, Spain, [2]Esplugues de Llobregat, Institut de Recerca Pediàtrica Hospital Sant Joan de Déu, Barcelona, Spain, [3]Mental Health Center Sct. Hans, 4000 Roskilde, Denmark and [4]Department of Family Medicine and Public Health, Unive California, San Diego, La Jolla, CA, USA

*To whom correspondence should be addressed

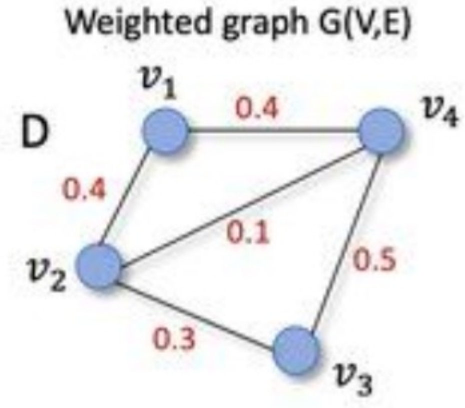
Associate Editor: Jonathan Wren
Received on January 21, 2020; revised on September 18, 2020; editorial decision on October 1, 2020; accepted on October 7, 2020

Heat diffusion on a network

v_i temperature of node *i*

$H_{i,j}$ thermal conductivity of edge *i* to *j*
(adjacency matrix)

l_i loss constant for node *i*



H

	v_1	v_2	v_3	v_4
v_1	0	0.4	0	0.4
v_2	0.4	0	0.3	0.1
v_3	0	0.3	0	0.5
v_4	0.4	0.1	0.5	0

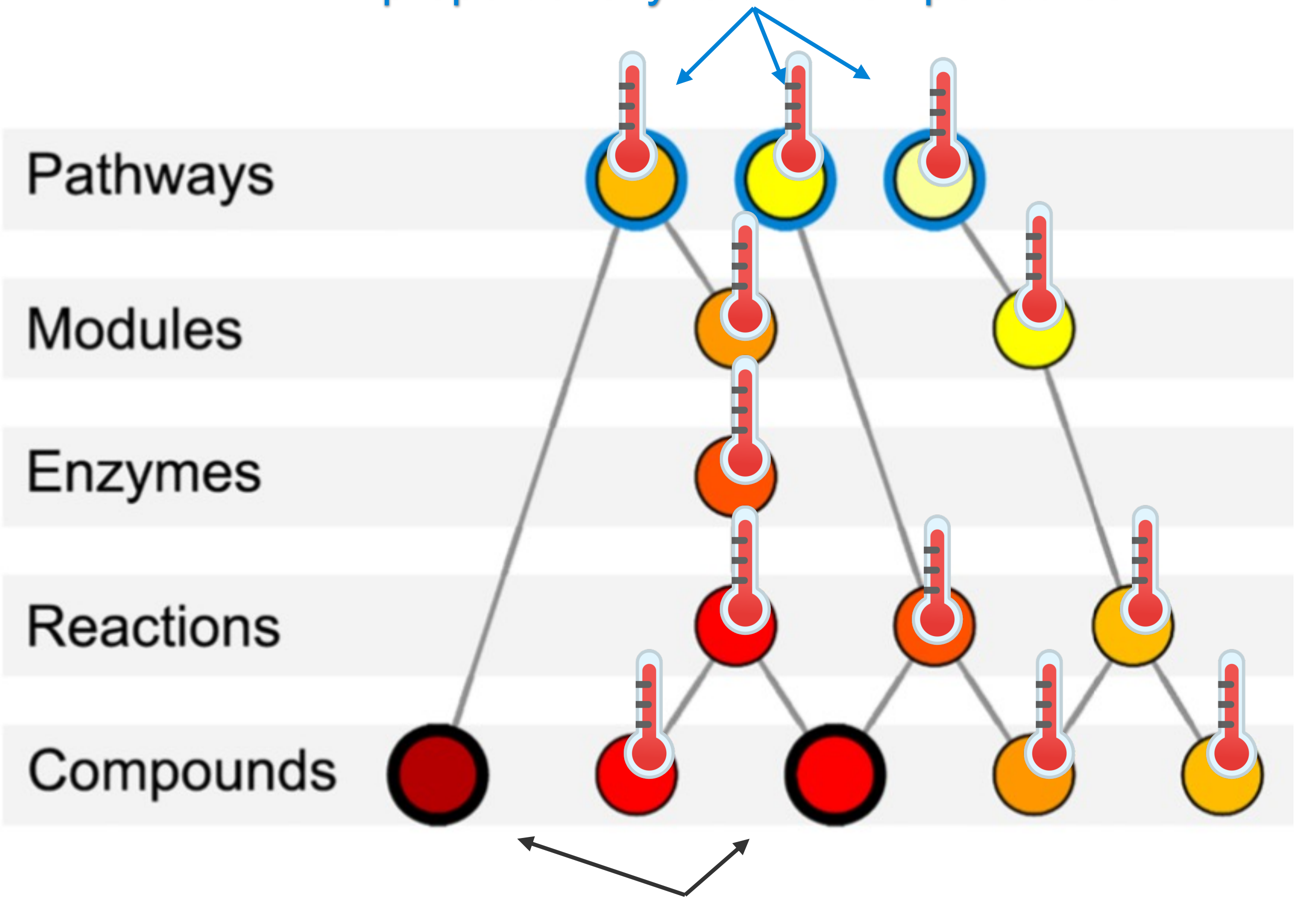
$$\frac{dv_i}{dt} = H_{i,j}\left(v_j - v_i\right) - l_i v_i$$

More about the mathematics of diffusion on a graph, including the definition of Laplacian at

`https://www.math.fsu.edu/~bertram/lectures/Diffusion.pdf`

FELLA principle

These blue nodes are dissipating heats proportionally to their temperatures



These black nodes belong to the observed metabolite universe . They are given a fixed temperature. For example, the metabolite that belong to a particular WGCNA module are given a temperature to 1° and all the others are set to 0°

Diffusion statistics principle

FELLA / DiffusStats comes with a fast method to compute means and standard deviations, in addition to the standard permutation simulation

Random perturbations of the fixed input temperatures. Diffusion to steady state and estimation the mean and standard deviation of temperature for every node in the network

Z-score (or another statistics)

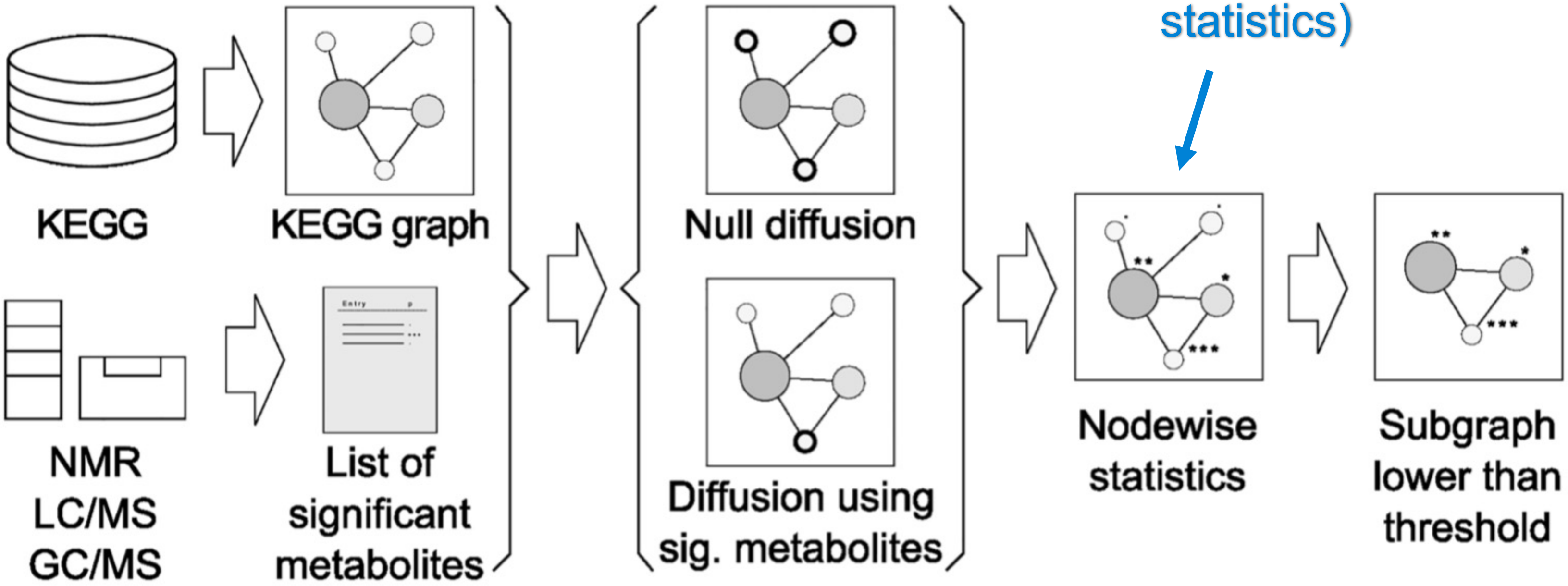
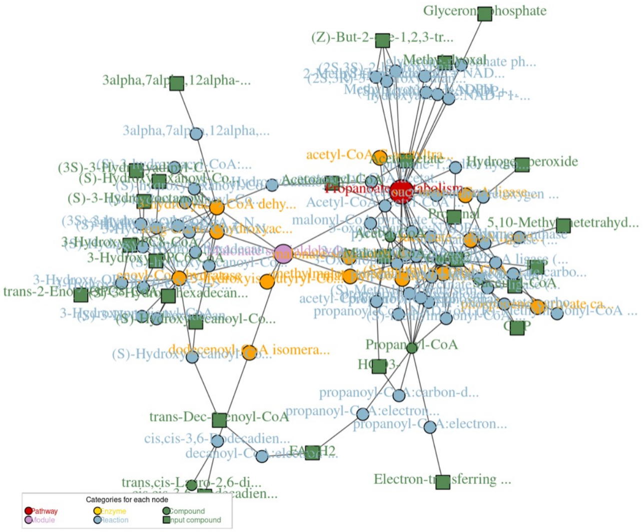

KEGG
KEGG graph
NMR
LC/MS
GC/MS
Entry
p
List of significant metabolites
Null diffusion
Diffusion using sig. metabolites
Nodewise statistics
Subgraph lower than threshold

Fig 1. Workflow summary. Contextual knowledge is extracted from KEGG as a graph object while experimental data is introduced as a list of affected metabolites. A null diffusive model assesses, and reports in a subgraph, which part of the KEGG graph is relevant for the input metabolites.

https://doi.org/10.1371/journal.pone.0189012.g001

What FELLA results looks like?

KEGG.id	Entry.type	KEGG.name	p.score
hsa00640	pathway	Propanoate metabolism - Homo sapiens (human)	0.0036894
M00013	module	Malonate semialdehyde pathway, propanoyl-CoA …	0.0044683
1.1.1.211	enzyme	long-chain-3-hydroxyacyl-CoA dehydrogenase	0.0371099
1.1.1.35	enzyme	3-hydroxyacyl-CoA dehydrogenase	0.0392511
1.2.1.18	enzyme	malonate-semialdehyde dehydrogenase (acetylat…	0.0069255
1.2.1.27	enzyme	methylmalonate-semialdehyde dehydrogenase (Co…	0.0165439
2.3.1.9	enzyme	acetyl-CoA C-acetyltransferase	0.0085923
3.1.2.4	enzyme	3-hydroxyisobutyryl-CoA hydrolase	0.0786804
4.1.1.32	enzyme	phosphoenolpyruvate carboxykinase (GTP)	0.0700429
4.1.1.41	enzyme	(S)-methylmalonyl-CoA decarboxylase	0.0223899
4.1.1.9	enzyme	malonyl-CoA decarboxylase	0.0002538
4.2.1.17	enzyme	enoyl-CoA hydratase	0.0015731
5.3.3.8	enzyme	dodecenoyl-CoA isomerase	0.0164255
6.2.1.4	enzyme	succinate—CoA ligase (GDP-forming)	0.0019142
6.2.1.5	enzyme	succinate—CoA ligase (ADP-forming)	0.0125330
R00209	reaction	pyruvate:NAD+ 2-oxidoreductase (CoA-acetylati…	0.0885938
R00233	reaction	malonyl-CoA carboxy-lyase (acetyl-CoA-forming…	0.0000698
R00238	reaction	Acetyl-CoA:acetyl-CoA C-acetyltransferase	0.0001037
R00353	reaction	malonyl-CoA:pyruvate carboxytransferase	0.0065794
R00405	reaction	Succinate:CoA ligase (ADP-forming)	0.0468613



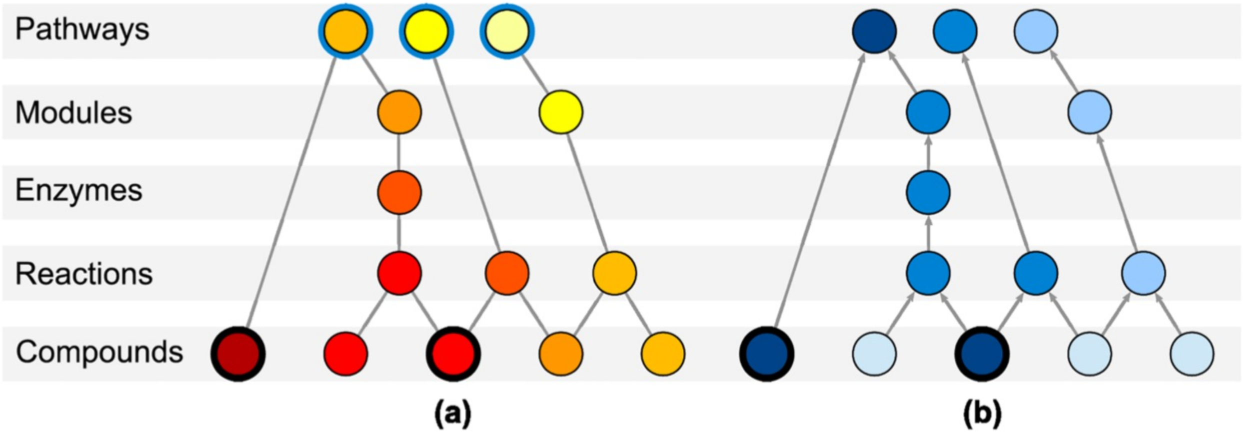


Fig 2. Nodes arrangement for (a) heat diffusion and (b) PageRank. The affected metabolites are highlighted with a black ring. For heat diffusion **(a)**, affected metabolites are forced to generate unitary flow. Every pathway is highlighted with a blue ring, representing its connection to a cool boundary node. In equilibrium, the highest temperature pathways (and nodes) will have the greatest heat flow, suggesting a relevant role in the experiment. For PageRank **(b)**, affected metabolites are the start of random walks. PageRank scores, represented by the intensity of the blue colour, will attain higher values in the frequently reached random walk nodes.

https://doi.org/10.1371/journal.pone.0189012.g002

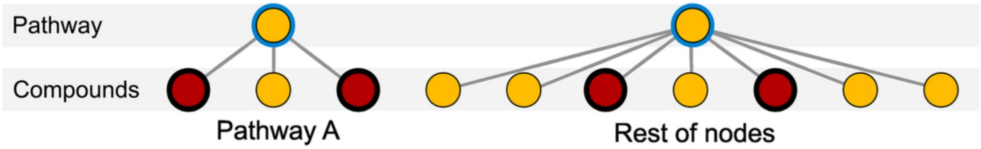


Fig 3. Toy example of an over-representation analysis of a hypothetical "pathway A" containing 3 metabolites out of a total of 10. The list to be enriched contains 4 metabolites, showing 2 hits in the pathway. The corresponding (Fisher's exact test) over-representation can be understood as a diffusion process on the depicted network followed by a null model. The temperature of pathway A is always coincident with the number of hits in the pathway, implying that its null distribution is the hypergeometric distribution, to which a one-tailed temperature comparison is made.

https://doi.org/10.1371/journal.pone.0189012.g003

Applications / limitations

Diffusion statistics advantages:

- works on (weighted) undirected network of any topology
- scales easily up to 20'000 nodes and any number of edges
- diffuStats implementation of Z-score computation run fast with a single set of observations

Limitations:

- only one type of edge
- no directionality or logical constraint can be expressed
- multiple sets of observations can possibly be investigated with the much slower monte-Carlo algorithm (I have not yet tested it)