A Model for Fine-Grained Data Citation

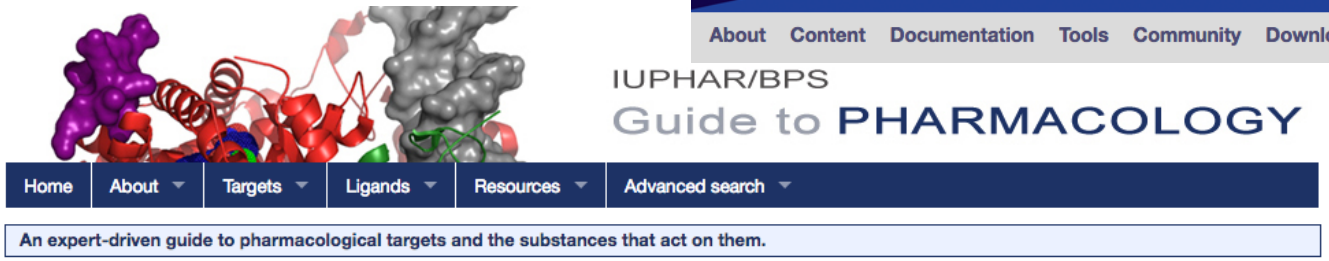
Susan B. Davidson, Daniel Deutch, Tova Milo, Gianmaria Silvello

Work partially supported by
NSF IIS 1302212, NSF ACI 1547360
NIH 3-U01-EB-020954-02S1
FP7 ERC grant MoDaS, agreement 291071
Israeli Science Foundation 1636/13
the Blavatnik Interdisciplinary Cyber Research Center.

National Science Foundation
WHERE DISCOVERIES BEGIN

Publication is changing

- Information is increasing published on the web.
- Much of this information is in **curated databases** – a mixture of crowd- or expert-sourced data and conventional publication.
- These datasets are complex, structured, and evolving, and contributors need to be acknowledged

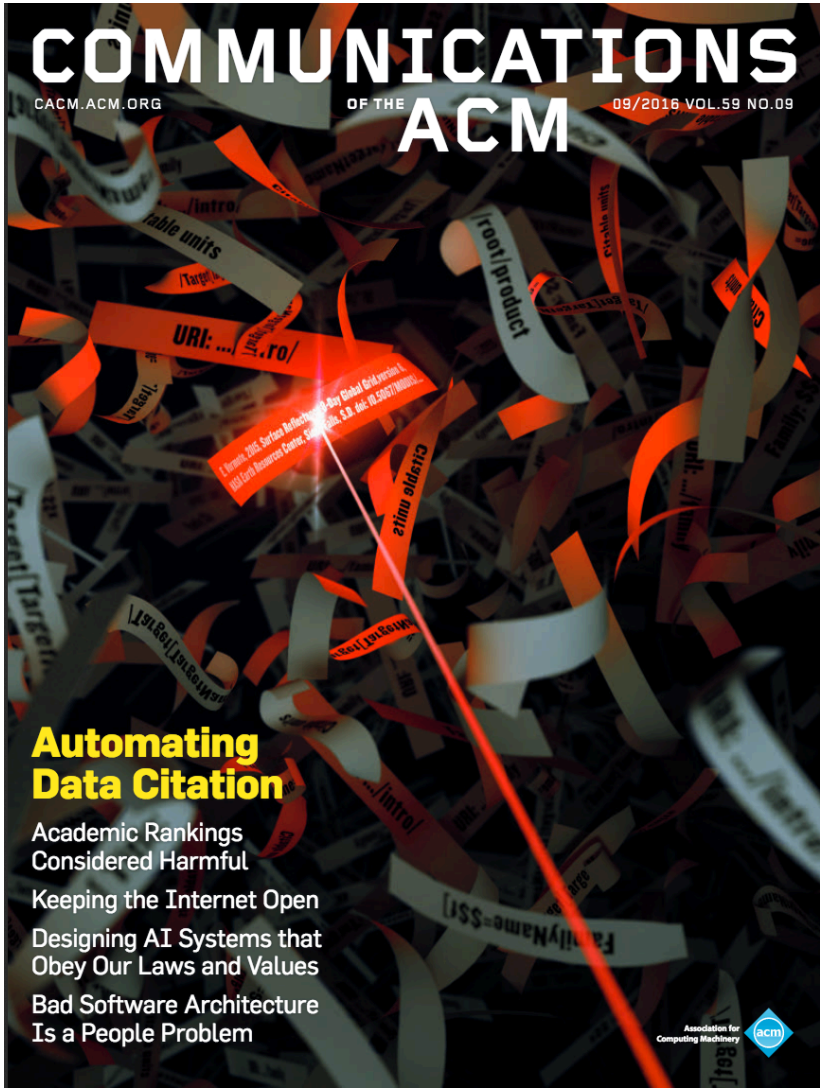


Increasing demand for data citation

- Large n… …ved: DataCi… …D-Lib Allianc… …MI, CODAT…
  - **Amste…** …sidered citable…
- Standa…
  - E.g Do…

```
<?xml version="1.0" encoding="UTF-8"?>
<!-- Revision history
    2010-08-26   Complete revision according to new common specification by the
                 metadata work group after review. AJH, DTIC
    2010-11-17 Revised to current state of kernel review, FZ, TIB
    2011-01-17 Complete revsion after community review. FZ, TIB
    2011-03-17 Release of v2.1: added a namespace; mandatory properties got
               minLength; changes in the definitions of relationTypes
    IsDocumentedBy/Documents and isCompiledBy/Compiles; changes type of
       property "Date" from xs:date to xs:string. FZ, TIB
    2011-06-27 v2.2: namespace: kernel-2.2, additions to controlled lists "resourceType",
"contributorType", "relatedIdentifierType", and "descriptionType". Removal of intermediate
include-files.
    2013-05 v3.0: namespace: kernel-3.0; delete LastMetadataUpdate &
MetadateVersionNumber; additions to controlled lists "contributorType",
"dateType", "descriptionType", "relationType", "relatedIdentifierType" &
"resourceType"; deletion of "StartDate" & "EndDate" from list "dateType"
and "Film" from "resourceType";  allow arbitrary order of elements; allow
optional wrapper elements to be empty; include xml:lang attribute for title, subject &
description; include attribute schemeURI for nameIdentifier of creator, contributor & subject;
added new attributes "relatedMetadataScheme", "schemeURI" & "schemeType" to
relatedIdentifier; included new property "geoLocation"
    2014-08-20 v3.1: additions to controlled lists "relationType", contributorType"
and "relatedIdentifierType"; introduction of new child element "affiliation" to
"creator" and "contributor"-->
<xs:schema xmlns:xs="http://www.w3.org/2001/XMLSchema" xmlns="http://
datacite.org/schema/kernel-3" targetNamespace="http://datacite.org/schema/
kernel-3" elementFormDefault="qualified" xml:lang="EN">
    <xs:import namespace="http://www.w3.org/XML/1998/namespace" s
chemaLocation="http://www.w3.org/2009/01/xml.xsd"/>
    <xs:include schemaLocation="include/datacite-titleType-v3.xsd"/>
    <xs:include schemaLocation="include/datacite-contributorType-v3.1.xsd"/>
    <xs:include schemaLocation="include/datacite-dateType-v3.xsd"/>
    <xs:include schemaLocation="include/datacite-resourceType-v3.xsd"/>
    <xs:include schemaLocation="include/datacite-relationType-v3.1.xsd"/>
    <xs:include schemaLocation="include/datacite-relatedIdentifierType-v3.1.xsd"/>
    <xs:include schemaLocation="include/datacite-descriptionType-v3.xsd"/>
    <xs:element name="resource">
```

Our manifesto…



- Principles and standards for data citation are unlikely to be used unless the process of extracting information is coupled with that of providing a citation for it.
- We need to **automatically generate citations** as the data is extracted.
- Data citation is a *computational* problem.

Buneman, Davidson, Frew:
Why data citation is a computational problem.
Commun. ACM 59(9): 50-57 (2016)

Outline

- **State of the art**
- Model: Citation views
- Citation “semi-rings”

What is a (conventional) citation?

- A collection of "snippets" of information: authors, title, date, etc. and some kind of access mechanism (DOI, URL, ISBN, shelf number etc.)
- Needed for a variety of reasons: kudos, currency, authority, recognition, access…
- Not exactly provenance

Cesar Palomo, Zhan Guo, Cláudio T. Silva, Juliana Freire:
Visually Exploring Transportation Schedules.
IEEE Trans. Vis. Comput. Graph. 22(1): 170-179 (2016)

Example 1: Eagle-I

- A "resource discovery" tool built to facilitate translational science research. Allows researchers to collect and share information about research resources (Core Facilities, iPS cell lines, software resources).
- Developed by a consortium of universities under NIH funding, headed by Harvard.
  - Penn is a member.
- Data is stored and distributed as RDF files (graph database).
- Resources have "Cite this resource" buttons!

Back to Search Results

Significance Tester for the Accumulation of Reads

Algorithmic software component

Send message to resource contact

Cite this resource

University of Pennsylvania

Software Description	STAR was developed to identify regions enriched for a histone modification based on ChIP-Seq evidence, by identifying regions with a significant accumulation of reads.
Software Additional Name	STAR
Used by	Computational Biology and Informatics Laboratory
Contact	Grant, Gregory R., Ph.D.
Related Technique	ChIP-seq assay
Software purpose	DNA modification site prediction objective

Back to Search Results

Significance Tester for the Accumulation of Reads

Algorithmic software component

Send message to resource contact

Cite this resource

University of Pennsylvania

eagle-i ID for this resource:

http://eagle-i.itmat.upenn.edu/i/0000013d-8d96-57e1-2ed7-105480000000

Click here for citation examples and more information.

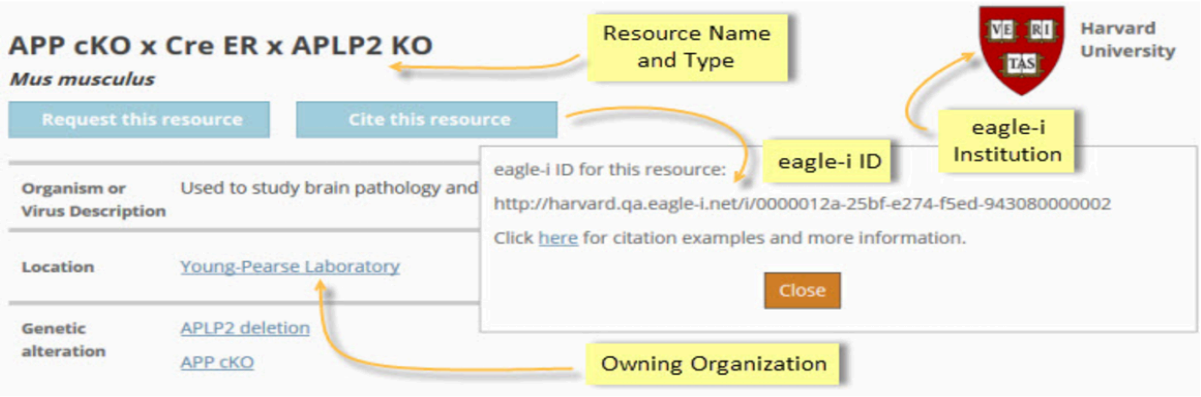
Close

Contact	Grant, Gregory R., Ph.D.
Related Technique	ChIP-seq assay

Citing an eagle-i Resource

Citing eagle-i resources is an easy way to give credit.

The formats suggested below provide the minimum information necessary to identify and credit the resource provider, and are designed to provide a traceable, durable, and unambiguous reference for the resource being cited. These suggestions can and should be used along with those from other resource identifiers (i.e. Antibody Registry ID, Addgene, DSHB, RRID) or from the journal publishing your work.



Note that for all types, the names of Core Facilities or other ambiguously named organizations should be followed by the name of the affiliated eagle-i institution in order to disambiguate them (e.g. *Flow Cytometry Core. Montana State University vs. Flow Cytometry Core. Dartmouth College*).

Citation Guidelines

Although only the most commonly cited types are listed below, the same rules can be used to cite any eagle-i resource.

Example 2: Reactome

About Content Documentation Tools Community Download Contact e.g. O95631, NTN1, signalin Search

About

About

Reactome is a free, open-source, curated and peer reviewed pathway database. Our goal is to provide intuitive bioinformatics tools for the visualization, interpretation and analysis of pathway knowledge to support basic research, genome analysis, modeling, systems biology and education.

About Reactome provides a background of the Reactome website, tools and research projects.

Reactome Team is a multi-disciplinary group of curators and software developers, located at the Ontario Institute for Cancer Research (OICR), New York University Medical Centre (NYUMC) and the European Bioinformatics Institute (EBI). Our aim is to identify important challenges in pathway curation analysis and vizualisation, and pursue high quality research while addressing those challenges.

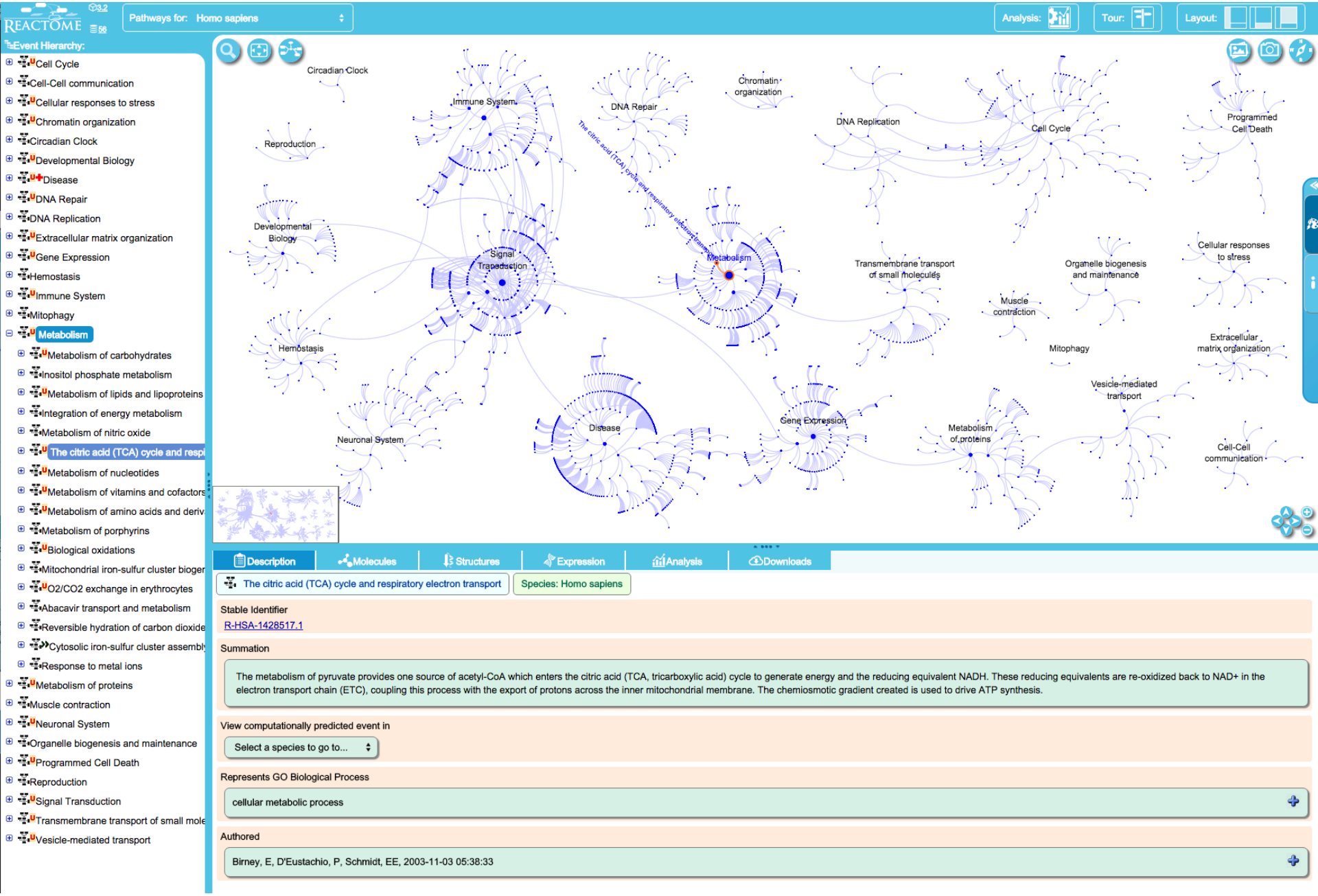
Scientific Advisory Board provides the Reactome group with independent, expert, multi-disciplinary, and strategic advice on our scientific research, emerging issues and trends, and on scientific partnerships and linkages.

Other Reactomes lists all the other Reactome databases that have been developed through a number of partnerships with other research groups and institutes worldwide.

License Agreement outlines the terms of the Creative Commons license.

About

- About Reactome
- News
- Reactome Team
- Scientific Advisory Board
- Other Reactomes
- License Agreement
- Reactome Disclaimer


REACTOME
Pathways for: Homo sapiens
Analysis:
Tour:
Layout:
Event Hierarchy:
Cell Cycle
Cell-Cell communication
Cellular responses to stress
Chromatin organization
Circadian Clock
Developmental Biology
Disease
DNA Repair
DNA Replication
Extracellular matrix organization
Gene Expression
Hemostasis
Immune System
Mitophagy
Metabolism
Metabolism of carbohydrates
Inositol phosphate metabolism
Metabolism of lipids and lipoproteins
Integration of energy metabolism
Metabolism of nitric oxide
The citric acid (TCA) cycle and respi
Metabolism of nucleotides
Metabolism of vitamins and cofactors
Metabolism of amino acids and deriv
Metabolism of porphyrins
Biological oxidations
Mitochondrial iron-sulfur cluster bioger
O2/CO2 exchange in erythrocytes
Abacavir transport and metabolism
Reversible hydration of carbon dioxide
Cytosolic iron-sulfur cluster assembly
Response to metal ions
Metabolism of proteins
Muscle contraction
Neuronal System
Organelle biogenesis and maintenance
Programmed Cell Death
Reproduction
Signal Transduction
Transmembrane transport of small mole
Vesicle-mediated transport
Circadian Clock
Immune System
DNA Repair
Chromatin organization
DNA Replication
Cell Cycle
Programmed Cell Death
Reproduction
The citric acid (TCA) cycle and respiratory electron transport
Developmental Biology
Signal Transduction
Metabolism
Transmembrane transport of small molecules
Organelle biogenesis and maintenance
Cellular responses to stress
Muscle contraction
Hemostasis
Mitophagy
Extracellular matrix organization
Vesicle-mediated transport
Disease
Gene Expression
Metabolism of proteins
Neuronal System
Cell-Cell communication
Description
Molecules
Structures
Expression
Analysis
Downloads
The citric acid (TCA) cycle and respiratory electron transport
Species: Homo sapiens
Stable Identifier
R-HSA-1428517.1
Summation
The metabolism of pyruvate provides one source of acetyl-CoA which enters the citric acid (TCA, tricarboxylic acid) cycle to generate energy and the reducing equivalent NADH. These reducing equivalents are re-oxidized back to NAD+ in the electron transport chain (ETC), coupling this process with the export of protons across the inner mitochondrial membrane. The chemiosmotic gradient created is used to drive ATP synthesis.
View computationally predicted event in
Select a species to go to...
Represents GO Biological Process
cellular metabolic process
Authored
Birney, E, D'Eustachio, P, Schmidt, EE, 2003-11-03 05:38:33

Referencing Reactome Publications

The following are examples that can be used to cite Reactome. To cite the Reactome project: If you are citing your use of Reactome in your work please cite these two recent Reactome publications:

- Fabregat et al. 2016 PMID: 26656494
- Milacic et al. 2012 PMID:24213504

To cite a Reactome pathway: Please use the appropriate DOI from the Table of Contents. You can read more about the way DOIs are assigned and used in Reactome here. You can add a DOI to the end of your citation following the appropriate style. Generally these citations follow this format: Author, A. (year). Title of article. Journal Title, X, xxx–xxx. doi:xxxxxx Please find examples here:

- APA Style
- Purdue U Online Writing Lab

To cite Reactome files obtained via the World Wide Web: Reactome project. "Reactome" http://www.reactome.org/ (date of message or visit).

To cite Reactome files available for download: Reactome project. "Reactome" http://www.reactome.org/download-data/ (date of access). When citing information obtained in a search of Reactome it should be remembered that while Reactome strives to contain the most current and accurate data, Reactome should not be used in citations where other primary sources of information are available.

Summary so far…

- Resources have some form of "persistent identifier"
  - Eagle-I gives it to you via "cite this resource" button
  - More complicated in Reactome
- Citations include the identifier and other more conventional snippets of information which is visible on the page but not provided automatically.
- Snippets of information to be included in the citation depend on the query.

Example 3: IUPHAR

- IUPHAR Guide to Pharmacology is a database of information about drug targets, and the prescription medicines and experimental drugs that act on them.
- Information is presented to users through a hierarchy of **web views**, with an **underlying relational implementation**.
- Contents of the database are generated by hundreds of experts who, in small groups, contribute to portions of the database. Thus the authorship depends on what part of the database is being cited.

Search database

IUPHAR/BPS Guide to PHARMACOLOGY

Home | About | Targets | Ligands | Resources | Advanced search

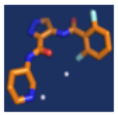
An expert-driven guide to pharmacological targets and the substances that act on them.

Targets

- G protein-coupled receptors
- Ion channels
- Nuclear hormone receptors
- Kinases
- Catalytic receptors
- Transporters
- Enzymes
- Other protein targets

Search for targets GO

Ligands



- Approved drugs
- Synthetic organics
- Metabolites
- Natural products
- Endogenous peptides
- Other peptides
- Inorganics
- Antibodies
- Labelled ligands

Search for ligands GO

What's new to Guide to PHARMACOLOGY

New version (2015.3) released 19th Oct 2015!

Target updates:

- GPCR updates:
  - Thyrotropin-releasing hormone receptors
  - NOP receptor
- VGIC updates:
  - Members of the Transient Receptor Potential superfamily of channels
- NHR updates:
  - Retinoic acid-related orphans introduction
  - Liver X receptor-α and β
  - COUP-TF-like receptors
  - 3-Ketosteroid receptors
- Enzyme updates:
  - Enzymes involved in hydrogen sulphide synthesis

BLAST search

Latest News

From our blog

GtoPdb database release 2015.3
by guidetopharmacology - Oct 19, 2015
Our total number of curated interactions now stands at 13859. New BLAST tool for sequence-based searching of targets available ...

New project to develop "The Guide to
by guidetopharmacology - Sep 25, 2015
We are very pleased to announce a new initiative (from 1st Nov 2015) to establish "The Guide to Immunopharmacology: ...

Latest news from NC-IUPHAR

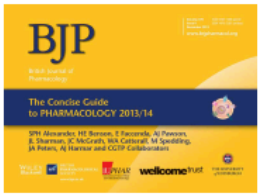
Hot topic: GPR3 may be a target for AD drug development
Oct 20, 2015
New paper describes how loss of GPR3 reduces the amyloid plaque burden and improves memory in Alzheimer's disease mouse models. ...

Database update: version 2015.3 released
Oct 20, 2015
The latest release includes new ligands, updates to GPCRs

Get email updates

Email address
Email format html text
Subscribe

The Concise Guide to PHARMACOLOGY 2013/14



A publication snapshot created from the database summary pages.

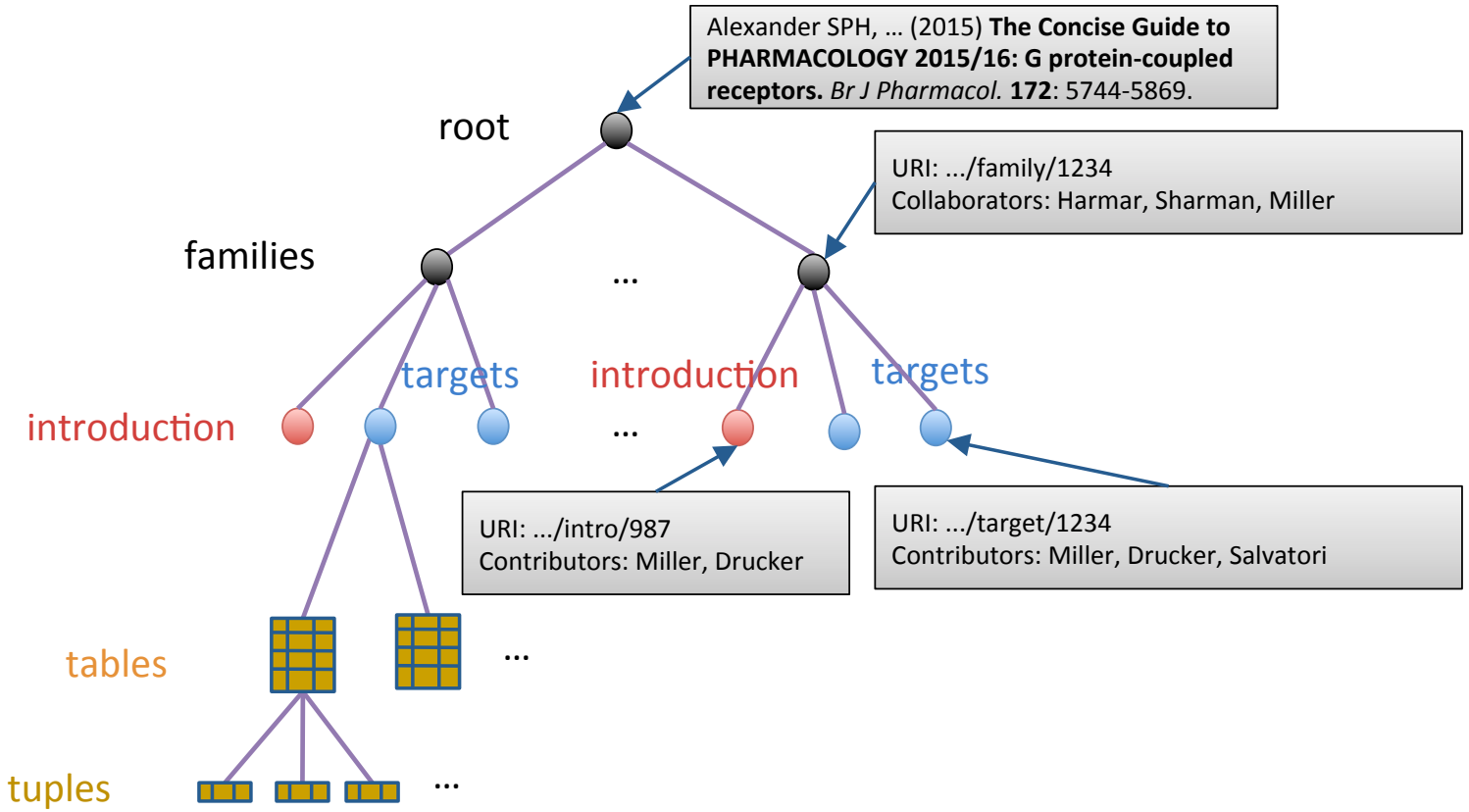
Access the table of contents GO

IUPHAR Database

IUPHAR DATABASE
International Union of Basic and Clinical Pharmacology

The IUPHAR/BPS Guide to PHARMACOLOGY builds upon and replaces the original IUPHAR Committee on Receptor Nomenclature and Drug Classification Database (IUPHAR-DB)

Citation structure in IUPHAR



Citations in IUPHAR

- Citations to objects retrieved via web pages are automatically generated in human readable form (embedded SQL)
- Want to lift these up to **schema-level "specifications"** of what the views are, how to obtain the citation snippets, and functions to display them in various forms (e.g. human readable, XML, BibTeX, RIS…)
- In the future, IUPHAR wants to enable citations to **general queries**

Why not just hard code citations?

- Citations vary with what part of of the database is being cited.
  - There are a very large number of "parts" of a database.
- A query may combine "parts" in intricate ways.
- We cannot expect to put a citation for each possible query result into DBLP.

Outline

- State of the art
- **Model: Citation views**
- Citation “semi-rings”

Returning to our manifesto

- The main problem:

  Given a database D and a query Q, generate an appropriate citation.

- **Database owners** need to be able to specify citations to parts of the database – schema level information.
- **Database users** need to have citations "served up" as they extract the data.
- **"Dereferencing"** the citation should bring back the data as of the time it was cited.

The citation generation problem

- It is common for the DBA to supply citations for some parts (**views**) of the database, $V_1 \ldots V_{n.}$.
- So the problem becomes: Given a query Q, can it be rewritten using the views? That is, is there a Q_i such that

$$\forall D \in S.\ Q(D) = Q_i(V_{i1}(D), \ldots, V_{ik}(D))$$

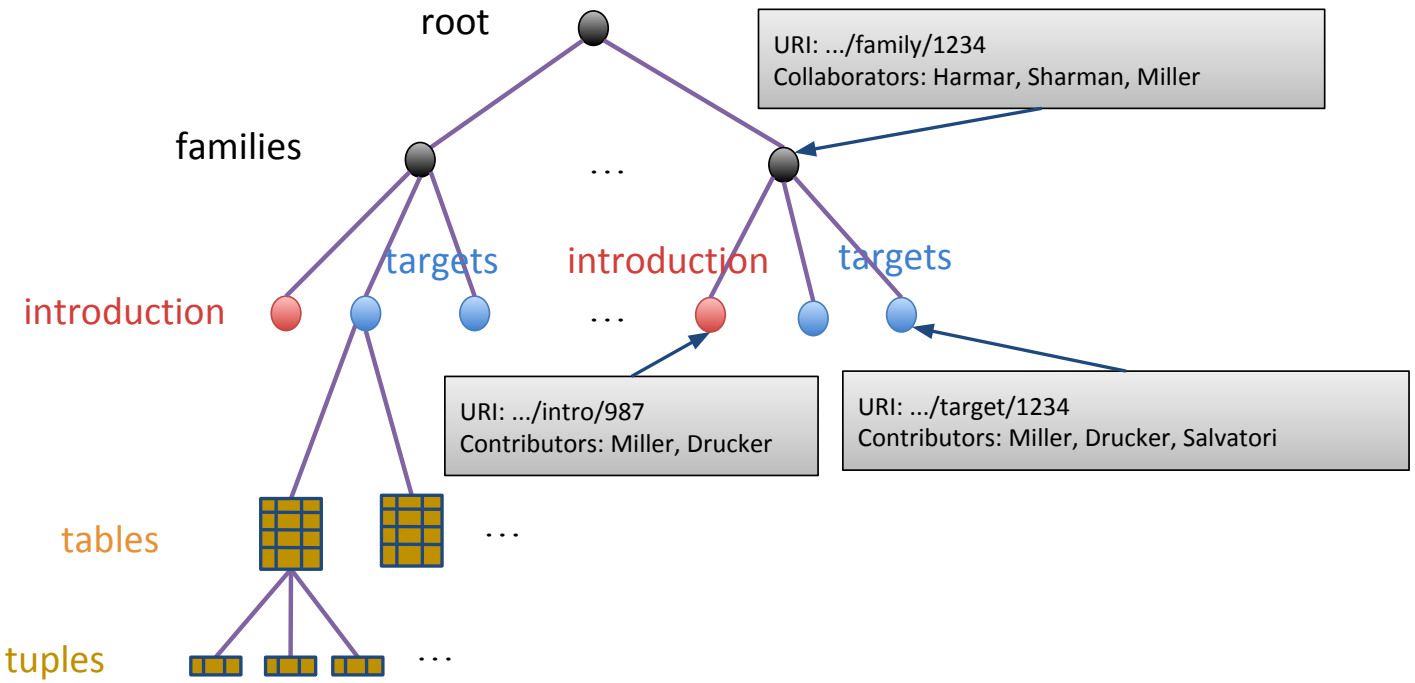
- If so, the citations for $V_{i1}\ldots, V_{ik}$ could be used to create (one or more) citations for Q.

Answering queries using views

- The problem of answering queries using views has been well studied and is generally hard – but in our context may be tractable.
  - A. Halevy. Answering queries using views: A survey. VLDB J., 10(4):270–294, 2001.
  - A. Deutsch, L. Popa, and V. Tannen. Query reformulation with constraints. SIGMOD Record, 35(1): 65–73, 2006.
  - F. Afrati, C. Li and J. Ullman. Using views to generate efficient evaluation plans for queries. JCSS 73(5): 703 - 724, 2007.

“Parameterized” views

- Owners may specify “parameterized” views
  - E.g. in IUPHAR there are views for family and family introduction pages, parameterized by FID, and views for target pages, parameterized by FID, TID



Citation views

- To specify a citation, there are three components:
  - ***View query***: specifies what is being cited
  - ***Citation query***: specifies what information to include in the ci
  - ***Citation function***: specifies how to construct the citation of information
- We call the triple a ***citation view***.
- What language(s) should we use?
  - For the view and citation query: Datalog
  - For the citation function: whatever you like!

"Universal" across different types of databases (e.g. relational, XML, RDF…)

Simplifies reasoning over queries and views

IUPHAR: Citation views

Schema:

Family(FID, FName, Type)
FamilyIntro(FID, Text)
Target(FID, TID, Info)
Person(PID, PName, Affiliation)
FC(FID, PID) FIC (FID, PID) FT(FID, TID, PID)

View queries:

λ F. V1(F, N,Ty) :- Family(F, N, Ty)
λ F. V2(F, Tx) :- FamilyIntro(F, Tx)
λ F, T. V3(F, T, I) :- Target (F, T, I)

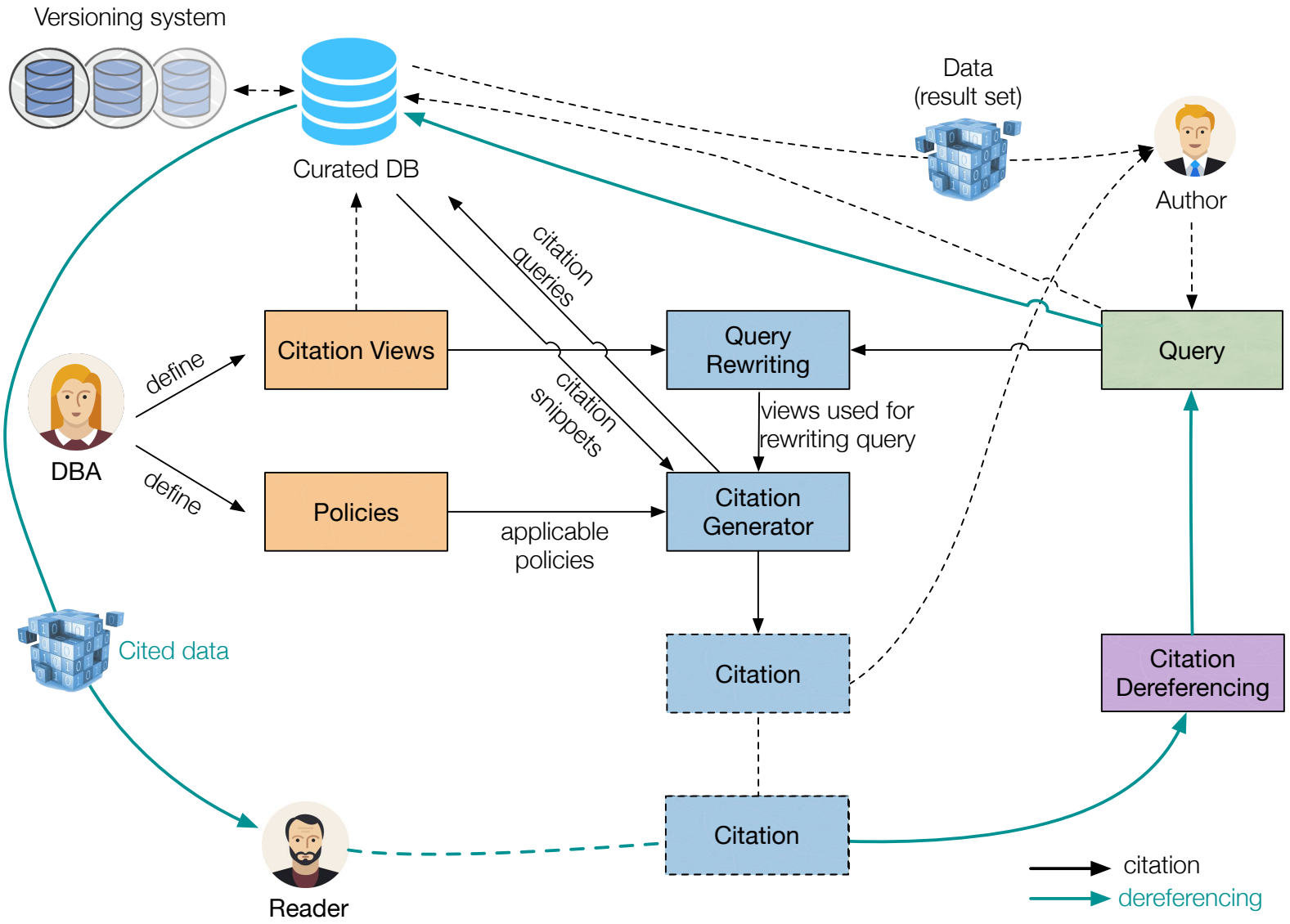
Citation queries:

λ F. C_{V1}(F, N, PN) :- Family(F, N, T), FC(F, P), Person(P, PN)
λ F. C_{V2}(F, N, PN) :- Family(F, N, Ty), FamilyIntro(F, Tx),
FIC(F, P), Person(P, PN)
λ F, T.C_{V3}(F, N, T, PN) :- Family(F, N, Ty), Target(F, T, I),
FT(F, T, P), Person(P, PN)

Generating citations

- If the query matches a view query, we can use the citation
  - "Match" must be extended to take parameters into account.
- But what if it doesn't?
  - Nothing matches the query
  - A set of view queries are used to rewrite the query
  - More than one set of view queries can be used to rewrite the query

Citation architecture



Outline

- State of the art
- Model: Citation views
- **Citation “semi-rings”**

Citations as annotation

- Citations are a type of annotation on tuples.
- Provenance is a form of annotation on tuples, which is well understood while being carried through queries.
- **Can we use these ideas to understand how citation "annotations" on tuples are combined in general queries?**

Citation "semi-ring"?

- Given a (conjunctive) query, we rewrite it to a set of minimal equivalent queries that contain at least one citation view.
  - Let the set of queries obtained in this way be $\{Q_1, ..., Q_n\}$
- Each Q_i contains a set of citation views $\{V_{i1}, ..., V_{imi}\}$. We use * to combine their citations to construct a citation for Q_i, $C(Q_i)$.
  - $C(Q_i) = C(V_{i1})*...*C(V_{imi})$
- C(Q) is constructed by + combining their citations.
  - $C(Q) = C(Q_1)+ ... + C(Q_n)$
  - E.g. + could be union or min (wrt some ordering on views)

Green, Karvounarakis, Tannen
PODS 2007: 31-40.

More on * and +

- **Joint** use of citations: $C(Q_i) = C(V_{i1})*...*C(V_{imi})$ could be union or some sort of join
  - E.g for spatio-temporal results, a minimal bounding box.
- **Alternate** use of citations: $C(Q) = C(Q_1)+ ... + C(Q_n)$
  - + could be union or min (wrt some ordering on views)
  - E.g. in IUPHAR, both the "Family" view and "Family Introduction" view are rewritings of a query on "Family Introduction", but "Family Introduction < Family"
- Joint and alternate use are "policies" specified by the DBA

Computational challenges

- Schema-level versus instance level?
  - Should we store the citations as annotations on tuples, or should we reason at the schema level and then calculate the citation?
- Given an expected query workload, what are the "best" citation views?
  - And are the necessary snippets of citation information in the schema?
- The number of rewritings of a given query is large.
  - Are there efficient algorithms to find the "best rewriting" according to some metric of quality (e.g. involving the number of views, the specificity of views, or related to a view hierarchy)?
- Scientometrics: measuring impact through citation views?

Conclusions

- If we want people to cite the data they use, we need to make it easy for them to do so.
- We must also make it easy for people who publish data to ensure that their data should be
- For many applications, citation of "parameterized views" for which citations can be attached.
- Joint and alternate use semantics are "policies" to be specified by the DBA

And there are many other interesting computational challenges with data citation!