Summer RAINET meeting

WP1: Data integration

Barcelona, July 2-3, 2015

Data integration : goals

- Identify all the data required for the study
- Determine their location (public DB) and how to get them
- Identify the important information in each source
- Build an analysis environment able to :
  - Integrate all the data
  - Allow to query any kind of relation/association between them
  - Allow to easily develop the future analysis
  - Associate the future analysis result to data to ensure reproducibility and durability
  - Allow easy share of results/information

Data Integration : identification

Proteins

- Protein definitions from Uniprot
- Protein cross-reference IDs
- Protein isoforms
- Protein domains

Gene Ontology

- Gene Ontology definition
- Gene ontology protein annotations

Kegg Pathway

- KEGG pathway definition
- Kegg Pathway protein annotations

Reactome pathway

- Reactome pathway definition
- Reactome pathway protein annotations

Interactome

- Detected protein interactions
- Partition of graph using OCG
- Functionnal annotation of partition modules

Data Integration : location and obtaining

wiki du projet tagc-rainet

Recent changes Media Manager Sitemap

Trace: • index • rainet-dataset

rainet-dataset

Table of Contents

Datasets for Work Package 1 (WP1)

Edit

Task 1.1

Edit

Proteomes

For the organisms of interest we use the proteomes available at http://www.uniprot.org. A UniProt proteome consists of the set of proteins thought to be expressed by an organism whose genome has been completely sequenced. Some proteomes have been (manually and algorithmically) selected as **reference proteomes**. They cover well-studied model organisms and other organisms of interest for biomedical research.

For each proteome we need: Sequences (FASTA files), IDs cross-references and Annotations.

FASTA files

```
ftp://ftp.uniprot.org/pub/databases/uniprot/current_release/knowledgebase/proteomes/DROME.fasta.gz
ftp://ftp.uniprot.org/pub/databases/uniprot/current_release/knowledgebase/proteomes/HUMAN.fasta.gz
ftp://ftp.uniprot.org/pub/databases/uniprot/current_release/knowledgebase/proteomes/YEAST.fasta.gz
```

IDs mapping files

```
ftp://ftp.uniprot.org/pub/databases/uniprot/current_release/knowledgebase/idmapping/by_organism/DROME_7227_idmap
```

Data Integration : location and obtaining

Proteins

- Protein definitions (Uniprot, 68561)
- Protein cross-reference Ids (Ensembl, 4315642)
- Protein isoforms (Uniprot, 38750)
- Protein domains (PFAM, 14881 ; SMART, 1274)

Gene Ontology

- Gene Ontology definition (GeneOntology, 43251)
- Gene ontology protein annotations (GeneOntology, 369769)

Kegg Pathway

- KEGG pathway definition (KEGG, 295)
- Kegg Pathway protein annotations (KEGG, 24825)

Reactome pathway

- Reactome pathway definition (Reactome, 23306)
- Reactome pathway protein annotations (Reactome, 316475)

Interactome

- Existing interactions (iMex compliant DB like DIP, IntAct, MINT...)
- Partition of graph using OCG (Brun group)
- Functionnal annotation of partition modules (Brun group)

Data Integration : location and obtaining

wiki du projet tagc-rainet

Recent changes Media Manager Sitemap

Trace: • index • rainet-dataset • **rainet-dataset-construction**

rainet-dataset-construction

Table of Contents

RAINET datasets construction

Edit

Proteome

Edit

Protocol

Source files (FASTA, Protein annotations, ID Mapping, Functionnal annotations) are listed in Datasets for Work Package 1 (WP1) (Drosophilia, human, yeast).

- Step 1: get the files from the database for a certain chosen fixed revision
- Step 2 : backup the files to a secure place to keep reference
- Step 3 : verify the degree of sequence redundancy in the fasta files. Check with Barcelona Team if catRAPID libraries consider redundancy or not. If redundancy has to be reduced, the redundancy threshold should be introduced.
- Step 4 : insert data in DB. Note that sequences will not be inserted. Instead the corresponding fasta files will be kept and parsed when sequences will be required (catRAPID). Insertion requires to cross FASTA files, annotation files and ID mapping files to get full protein information in DB.

Note: Do we have to take into account isonforms of proteins? catRAPID classically uses proteome with only one canonical sequence per protein (UniprotKB) and compares them to the whole transcriptomes from Ensembl. Would it be interesting to take into account also the isoforms of proteins?

Note: for cross references, we have to be aware of the correspondance between the chosen version of UniprotKB and Ensembl.

Note: for GO annotations, annotation file and GO obo file must be inline (same revision).

Edit

Tables

Table Protein:

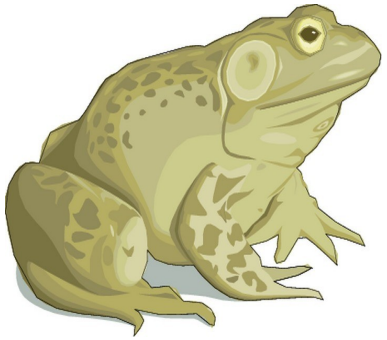
- The content of this table is generated by the query on UniprotKB (see UNIPROT QUERY on Datasets for Work Package 1 (WP1). Please note that the dbSource column is parsed from the the header of the proteome FASTA files.

Data integration : goals

- Identify all the data required for the study
- Determine their location (public DB) and how to get them
- Identify the important information in each source
- Build an analysis environment able to :
  - Integrate all the data
  - Allow to query any kind of relation/association between them
  - Allow to easily develop/execute the future analysis
  - Associate the future analysis result to data to ensure reproducibility and durability
  - Allow easy share of results/information

Data integration : constructing environment

Rainet2Toad

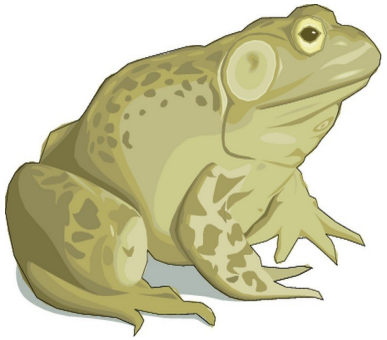


Total Omic Analysis of Data

A fully Object-Oriented Python software with complete ORM database integration

Data integration : constructing environment

Rainet2Toad



Total Omic Analysis of Data

A fully Object-Oriented Python software with complete ORM database integration

- Object-Oriented : easy to develop, easy to maintain, easy to evolve, robust model
- Python : language commonly known by bioinformaticians, strong bio-libraries
- Database (SQLite) : powerful query with SQL, data association, durability, reproducibility, easy to share (single file) and to restore
- ORM (SQLAlchemy) : easy DB management, easy DB query, easy object management

Data integration : code with standard development

Code uses standard design patterns

- Factory
- Singleton
- Strategy

Code uses standard error management

- Raising/catching of exceptions
- Multi-level logging protocol

Code uses standard documentation

- Internal code documentation
- Doxygene API documentation

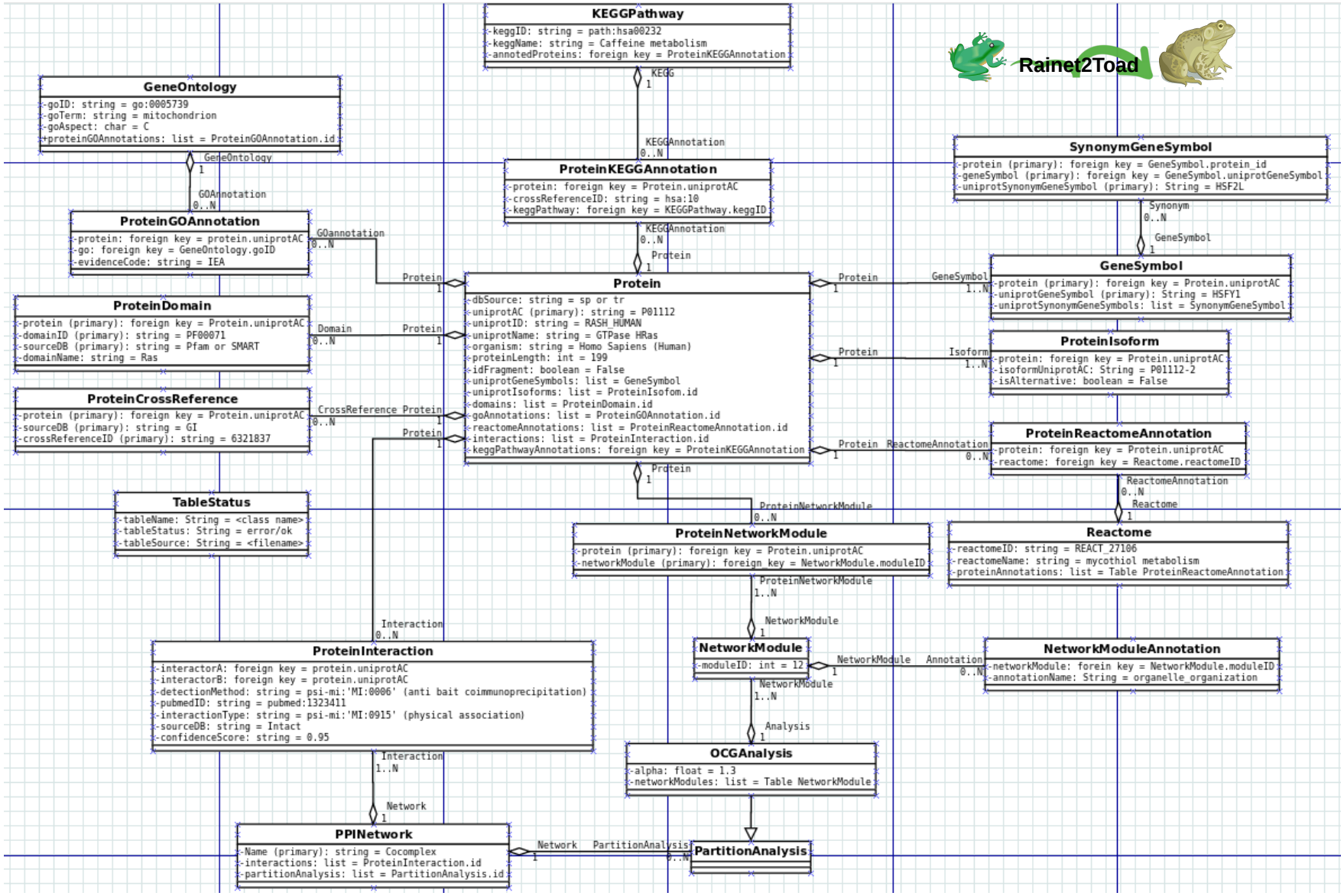
Code deployment is automated

- ant task used to deploy code on server
- ant task used to deploy/uncompress data on server

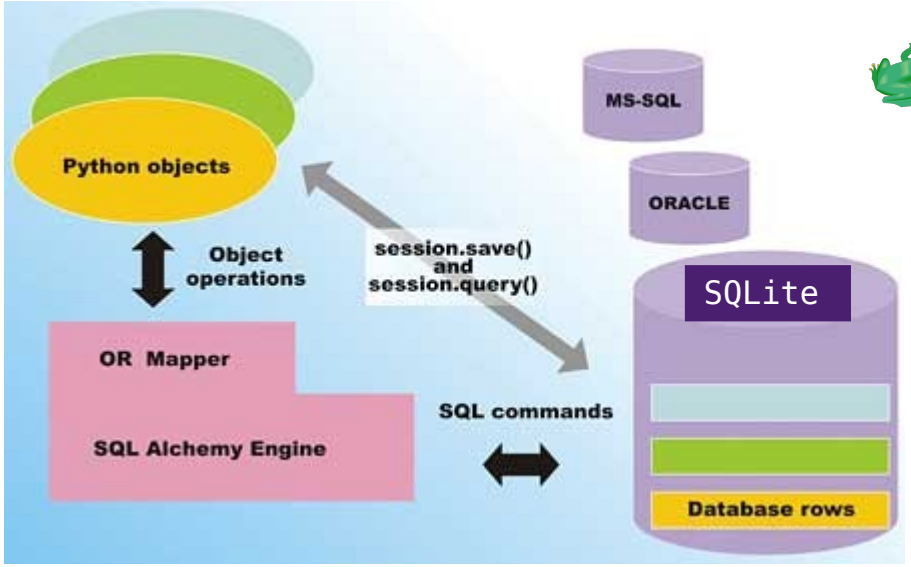
Execution is done through easy-to-use command line

- Command are similar to tools like samtools
- Documented through python standards

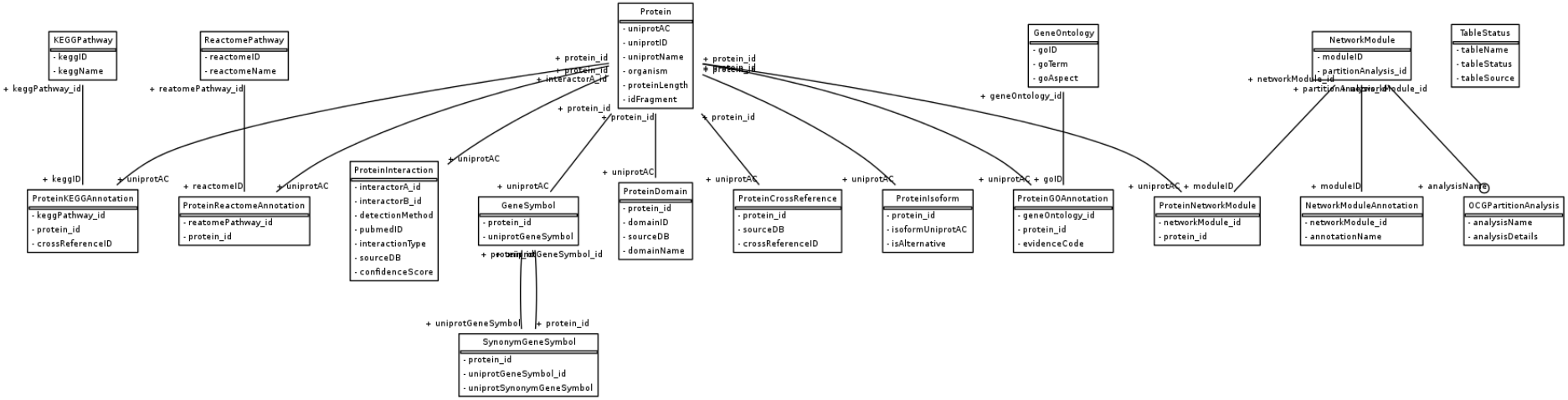
Data integration : class model



Data integration : ORM and database tables



Rainet2Toad



Data integration : easy to use DB Firefox plugin

Rainet2Toad

SQLite Manager - /tmp/Rainet/rainet.db.sqlite

Database Table Index View Trigger Tools Help

Directory (Select Profile Database) Go

rainet.db.sqlite

- Master Table (1)
- Tables (18)
  - GeneOntology
  - GeneSymbol
  - KEGGPathway
  - NetworkModule
  - NetworkModuleAnnotation
  - OCGPartitionAnalysis
  - Protein
  - ProteinCrossReference
  - ProteinDomain
  - ProteinGOAnnotation
  - ProteinInteraction
  - ProteinIsoform
  - ProteinKEGGAnnotation
  - ProteinNetworkModule
  - ProteinReactomeAnnotation
  - ReactomePathway
  - SynonymGeneSymbol
  - TableStatus
- Views (0)
- Indexes (18)
- Triggers (0)

Structure | Browse & Search | Execute SQL | DB Settings

TABLE Protein Search Show All Add Duplicate Edit Delete

rowid	uniprotAC	uniprotID	uniprotName	organism	proteinLength	idFragment
1	P31946	1433B_HUMAN	14-3-3 protein beta/alp...	Homo sapiens (Human)	246	0
2	P62258	1433E_HUMAN	14-3-3 protein epsilon (1...	Homo sapiens (Human)	255	0
3	Q04917	1433F_HUMAN	14-3-3 protein eta (Prot...	Homo sapiens (Human)	246	0
4	P61981	1433G_HUMAN	14-3-3 protein gamma (...	Homo sapiens (Human)	247	0
5	P31947	1433S_HUMAN	14-3-3 protein sigma (Ep...	Homo sapiens (Human)	248	0
6	P27348	1433T_HUMAN	14-3-3 protein theta (14-...	Homo sapiens (Human)	245	0
7	P63104	1433Z_HUMAN	14-3-3 protein zeta/delt...	Homo sapiens (Human)	245	0
8	P30443	1A01_HUMAN	HLA class I histocompat...	Homo sapiens (Human)	365	0
9	P01892	1A02_HUMAN	HLA class I histocompat...	Homo sapiens (Human)	365	0
10	P04439	1A03_HUMAN	HLA class I histocompat...	Homo sapiens (Human)	365	0
11	P13746	1A11_HUMAN	HLA class I histocompat...	Homo sapiens (Human)	365	0
12	Q96QU6	1A1L1_HUMAN	1-aminocyclopropane-1-...	Homo sapiens (Human)	501	0
13	Q4AC99	1A1L2_HUMAN	Probable inactive 1-ami...	Homo sapiens (Human)	568	0
14	P30447	1A23_HUMAN	HLA class I histocompat...	Homo sapiens (Human)	365	0
15	P05534	1A24_HUMAN	HLA class I histocompat...	Homo sapiens (Human)	365	0
16	P18462	1A25_HUMAN	HLA class I histocompat...	Homo sapiens (Human)	365	0
17	P30450	1A26_HUMAN	HLA class I histocompat...	Homo sapiens (Human)	365	0
18	P30512	1A29_HUMAN	HLA class I histocompat...	Homo sapiens (Human)	365	0
19	P16188	1A30_HUMAN	HLA class I histocompat...	Homo sapiens (Human)	365	0
20	P16189	1A31_HUMAN	HLA class I histocompat...	Homo sapiens (Human)	365	0
21	P10314	1A32_HUMAN	HLA class I histocompat...	Homo sapiens (Human)	365	0
22	P16190	1A33_HUMAN	HLA class I histocompat...	Homo sapiens (Human)	365	0
23	P30453	1A34_HUMAN	HLA class I histocompat...	Homo sapiens (Human)	365	0
24	P30455	1A36_HUMAN	HLA class I histocompat...	Homo sapiens (Human)	365	0
25	P30456	1A43_HUMAN	HLA class I histocompat...	Homo sapiens (Human)	365	0
26	P30457	1A66_HUMAN	HLA class I histocompat...	Homo sapiens (Human)	365	0
27	P01891	1A68_HUMAN	HLA class I histocompat...	Homo sapiens (Human)	365	0
28	P10316	1A69_HUMAN	HLA class I histocompat...	Homo sapiens (Human)	365	0
29	Q09160	1A80_HUMAN	HLA class I histocompat...	Homo sapiens (Human)	365	0
30	P30459	1A74_HUMAN	HLA class I histocompat...	Homo sapiens (Human)	365	0
31	P01889	1B07_HUMAN	HLA class I histocompat...	Homo sapiens (Human)	362	0
32	P30460	1B08_HUMAN	HLA class I histocompat...	Homo sapiens (Human)	362	0
33	P30461	1B13_HUMAN	HLA class I histocompat...	Homo sapiens (Human)	362	0
34	P30462	1B14_HUMAN	HLA class I histocompat...	Homo sapiens (Human)	362	0
35	P30464	1B15_HUMAN	HLA class I histocompat...	Homo sapiens (Human)	362	0
36	P30466	1B18_HUMAN	HLA class I histocompat...	Homo sapiens (Human)	362	0
37	P03989	1B27_HUMAN	HLA class I histocompat...	Homo sapiens (Human)	362	0

<< < 1 to 100 of 68561 > >>

SQLite 3.8.6 Gecko 35.0.1 0.8.3.1-signed Exclusive Number of files in selected directory: 7 ET: 8 ms

Data integration : next tasks

Rainet2Toad

Build interactome

- Retrieve the correct set of interactions
- Insert it to DB

Analyse interactome

- Build partition of graph
- Build functionnal annotations of partition modules

Develop first analysis