Protein Graph Database

Borna Bešić, Dilmurat Yusuf

Bioinformatics Group

Albert-Ludwigs-Universität, Freiburg

March 11, 2019

Overview

What do we want?

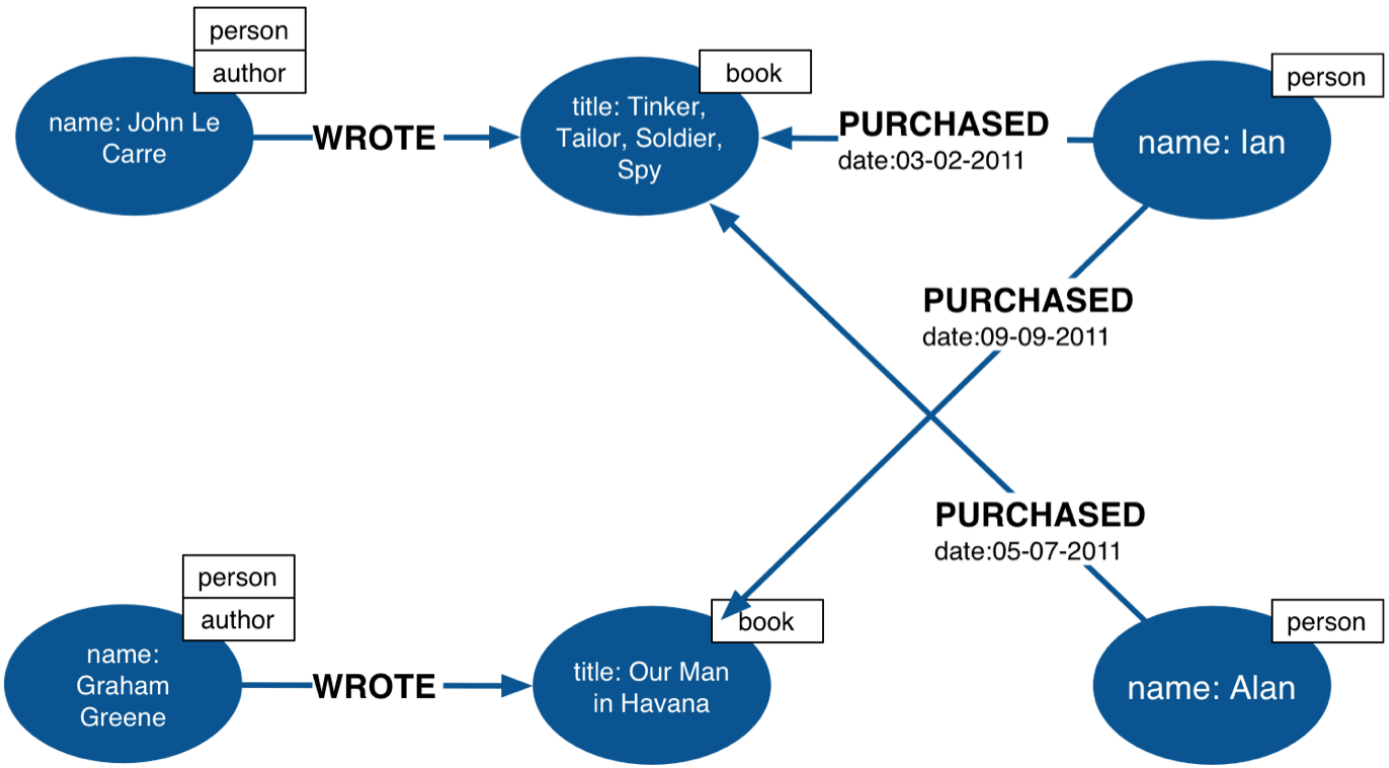
- Easier access to data
  - More flexible searches
  - Faster queries
  - Dynamic user interfaces
- Associated data in one place
  - Unification of databases
- This leads to
  - Accessible protein - disease data
  - Faster drug discovery

STRING

- Database of known and predicted protein-protein interactions
- 7 score channels $\rightarrow$ combined score
  1. Experiments
  2. Database
  3. Textmining
  4. Co-expression
  5. Neighbourhood
  6. Fusion
  7. Co-occurence
- Best overall performance for gene set recovery by propagation (Huang et al., Systematic Evaluation of Molecular Networks for Discovery of Disease Genes, Cell Systems)
- Full database dumps available for download: 512.8 GB (compressed)

Graph databases

- Nodes
- **Directed** edges / relationships
- Properties



Neo4j & Cypher

Example of a Cypher query

```
MATCH (:person { name: "Alan" })-[p:PURCHASED]->(b:book)
RETURN p.date, b
```

- Pros:
  - Outperforming RDBMS for associative data
  - No redundancy
  - Easier to change the design
- Cons:
  - Not standard; new query language (Cypher)
  - Harder to do summing queries and max queries efficiently

Why Neo4j?

	Neighbor network	Best-scoring path	Shortest path
PostgreSQL	206.31 s	1147.74 s	976.22 s
Neo4j	5.68 s[a]	1.17 s	0.40 s
Speedup	36×	981×	2441×

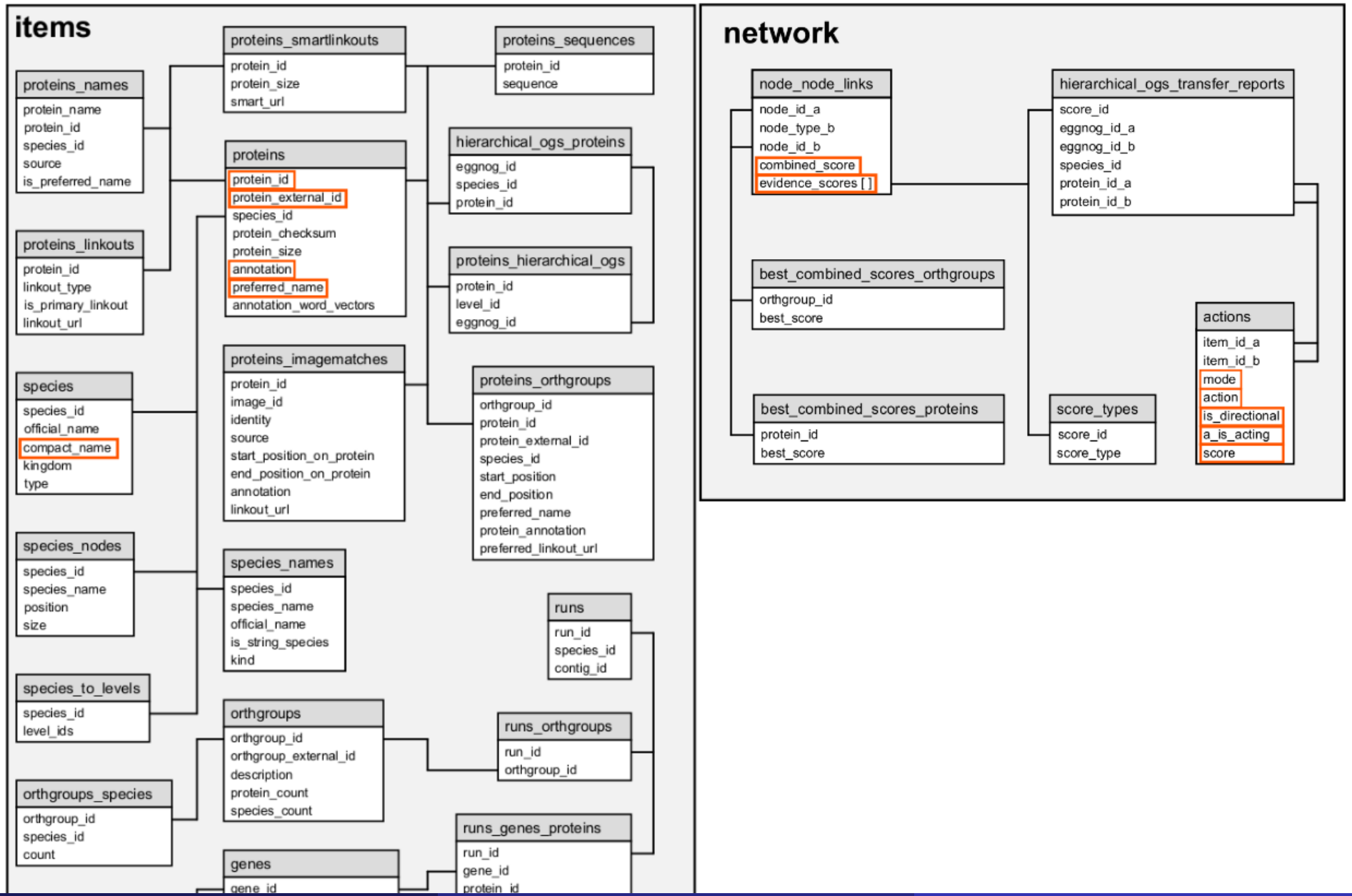
Figure: Have CT, Jensen LJ. Are graph databases ready for bioinformatics?

Overview

Sources of data

- STRING
  - proteins
  - protein - protein associations
- KEGG PATHWAY
  - pathways
    - classes
    - compounds
    - drugs
    - diseases

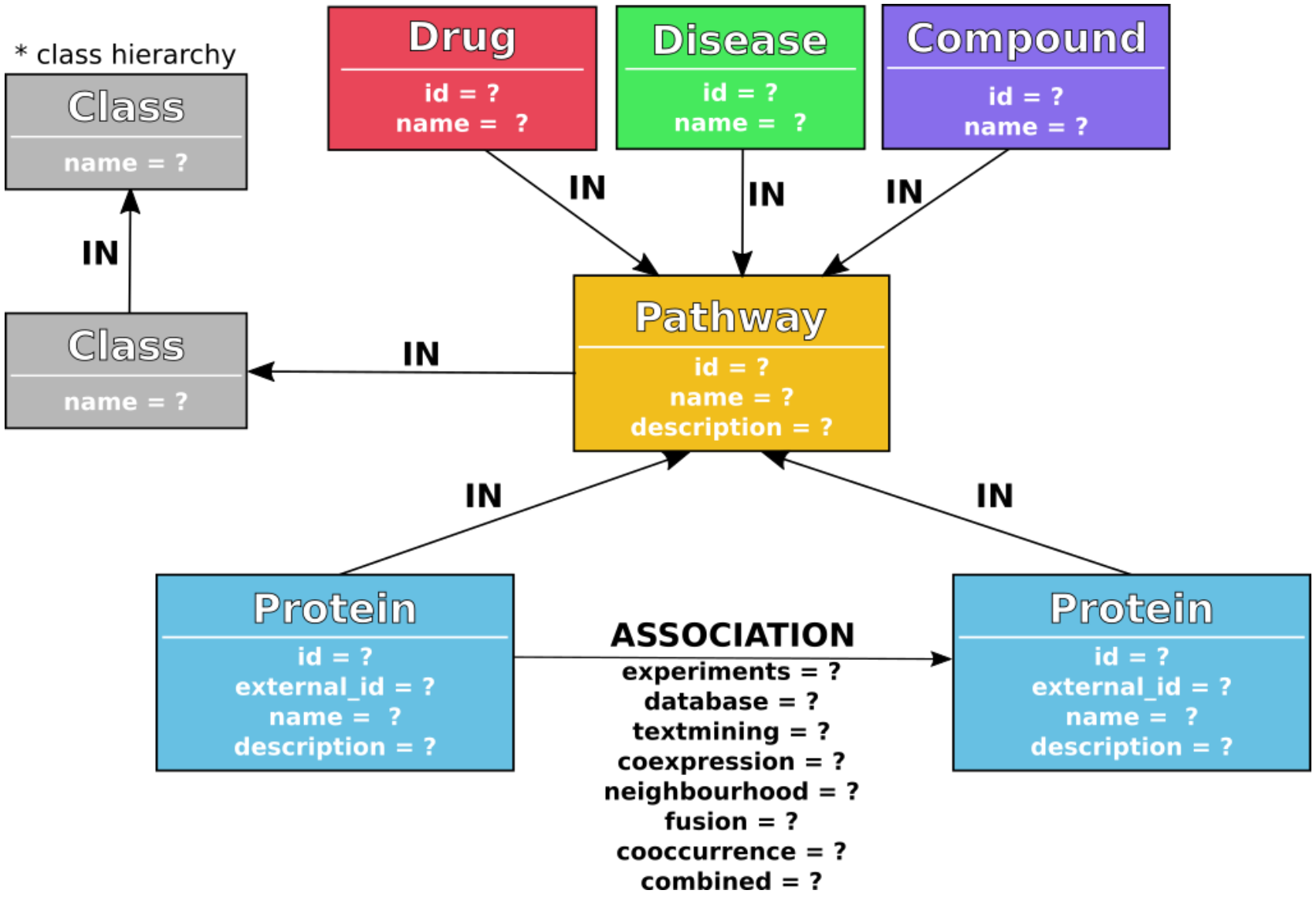
STRING Schema



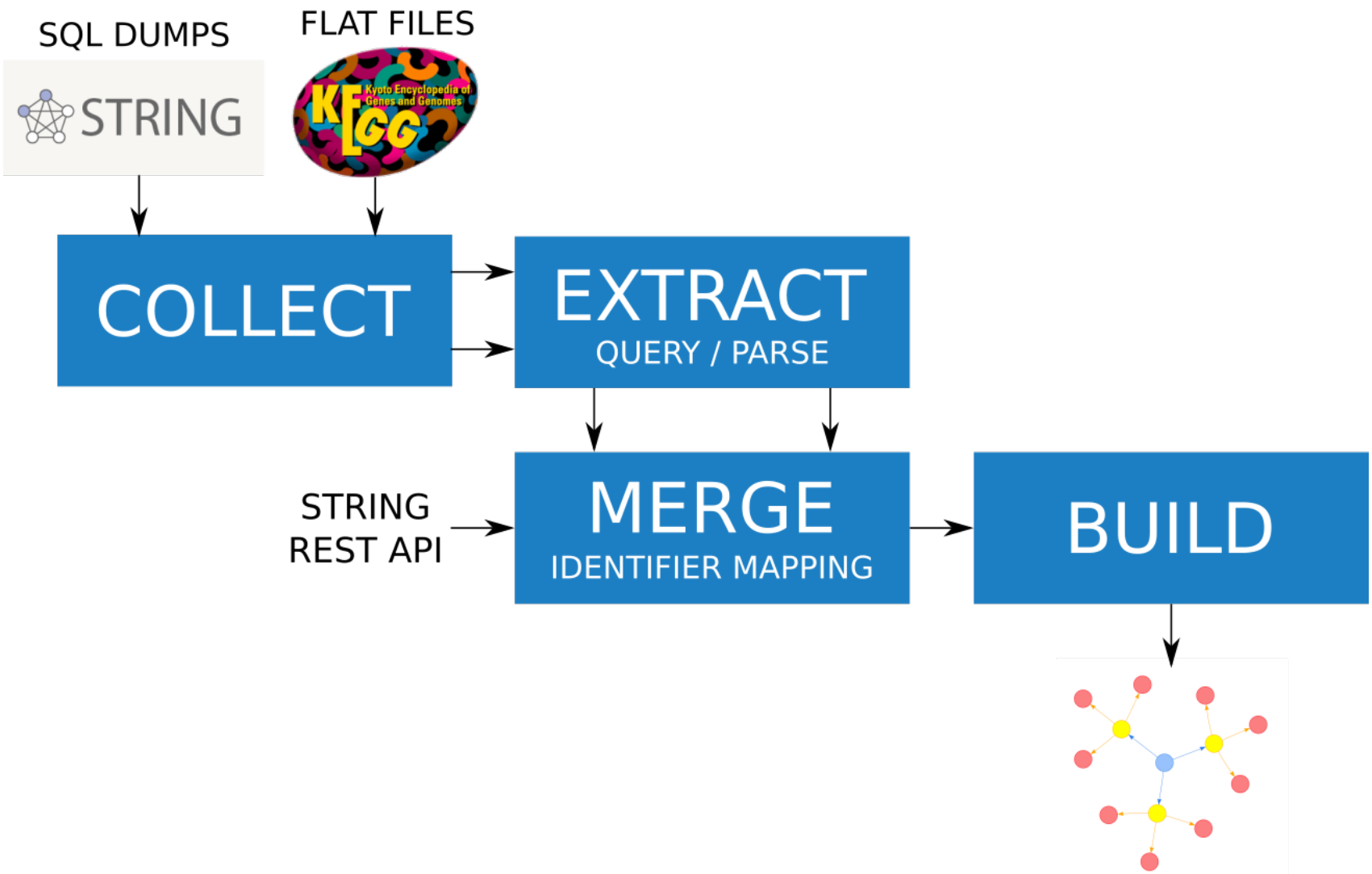
KEGG PATHWAY Flat file

```
ENTRY       hsa00010                    Pathway
NAME        Glycolysis / Gluconeogenesis - Homo sapiens (human)
DESCRIPTION Glycolysis is the process of converting glucose into pyruvate...
CLASS       Metabolism; Carbohydrate metabolism
PATHWAY_MAP hsa00010  Glycolysis / Gluconeogenesis
MODULE      hsa_M00001  Glycolysis (Embden-Meyerhof pathway), glucose => pyruvate [PATH:hsa00010]
            hsa_M00002  Glycolysis, core module involving three-carbon compounds [PATH:hsa00010]
            hsa_M00003  Gluconeogenesis, oxaloacetate => fructose-6P [PATH:hsa00010]
            hsa_M00307  Pyruvate oxidation, pyruvate => acetyl-CoA [PATH:hsa00010]
DISEASE     H00069  Glycogen storage disease
            H00071  Hereditary fructose intolerance
            H00072  Pyruvate dehydrogenase complex deficiency
            ...
DRUG        D00123  Cyanamide (JP17)
            D00131  Disulfiram (JP17/USP/INN)
            D07257  Lonidamine (INN)
            D08970  Piragliatin (USAN)
DBLINKS     BSID: 82926
            GO: 0006096 0006094
ORGANISM    Homo sapiens (human) [GN:hsa]
GENE        3101  HK3; hexokinase 3 [KO:K00844] [EC:2.7.1.1]
            3098  HK1; hexokinase 1 [KO:K00844] [EC:2.7.1.1]
            3099  HK2; hexokinase 2 [KO:K00844] [EC:2.7.1.1]
            80201  HKDC1; hexokinase domain containing 1 [KO:K00844] [EC:2.7.1.1]
            2645  GCK; glucokinase [KO:K12407] [EC:2.7.1.2]
            ...
COMPOUND    C00022  Pyruvate
            C00024  Acetyl-CoA
            C00031  D-Glucose
            C00033  Acetate
            ...
REFERENCE   (map 1)
  AUTHORS   Nishizuka Y (ed).
  TITLE     [Metabolic Maps] (In Japanese)
  JOURNAL   Tokyo Kagaku Dojin (1980)
REFERENCE   (map 1)
  AUTHORS   Nishizuka Y, Seyama Y, Ikai A, Ishimura Y, Kawaguchi A (eds).
  TITLE     [Cellular Functions and Metabolic Maps] (In Japanese)
  JOURNAL   Tokyo Kagaku Dojin (1997)
REFERENCE
  AUTHORS   Michal G.
  TITLE     Biochemical Pathways
  JOURNAL   Wiley (1999)
```

Protein Graph DB Scheme



Workflow


SQL DUMPS
FLAT FILES
STRING
KEGG
Kyoto Encyclopedia of Genes and Genomes
COLLECT
EXTRACT
QUERY / PARSE
STRING
REST API
MERGE
IDENTIFIER MAPPING
BUILD

Overview

Protein Graph DB in numbers

- Species: Homo sapiens (human) & Mus musculus (house mouse)
- Nodes
  - Protein: 43 125
  - Pathway: 646
  - Drug: 3 731
  - Disease: 969
  - Compound: 3 465
  - Class: 49
- Relationships
  - **ASSOCIATION**: 11 983 549
  - **IN**: 76 715

(demo)

Overview

Next steps

- Extend the database
  - Protein - protein actions
  - KEGG DRUG
  - KEGG DISEASE
  - KEGG COMPOUND
- Other species
- Production-level web server
- Machine learning

Thanks :)

https://backofenlab.github.io/protein-graph-database/
https://github.com/BackofenLab/protein-graph-database