Informationsintegration
Übung 3

SS 2016

Yvonne Lichtblau

Vorstellung Lösungen Übung 2

Assignment 3
Hierarchical Queries

Overview

- Reuse all data you have imported in Assignment 2
- In this task we will
  - Clean it up a bit
  - Integrate the real Gene Ontology
  - Answer queries that need to traverse a taxonomy

Task 1: Data Cleaning

- Eliminate all entries from your gene table (merged) which do not belong to humans (tax_id for humans is '9606')
- Duplicate elimination
  - The gene table (merged) has duplicates (multiple entries with the same geneid). These may have different status, protein accession, start/end positions etc.
  - Check for each attribute wheter it functionally depends on the geneid column
  - All values which have 1:n relationship to a gene should be moved to a separate table
    - One table for the essential data
    - Start and end position should be saved together
- Define a primary key on GENEID on your gene table

Task 2: Integrate the Gene Ontology (1)

- The Gene Ontology is a collection of 3 ontologies (GO:0003674 – molecular_function, GO:0008150 – biological_process, GO:0005575 – cellular_component)
  - Each forms a DAG
  - Semantics: If a gene G is assigned aterm X and X IS_A Y (directly or indirectly), then Y is implicitly assigned to G
- Model such an ontology in your schema
  - We want: term_id (go_id), term_name, ontology, IS_A relationships
  - Example:
    - term_id/go_id: GO:0032421
    - term_name: stereocilium bundle
    - ontology: cellular_component
    - IS_A relationships: GO:0097458, GO:0098862

Task 2: Integrate the Gene Ontology (2)

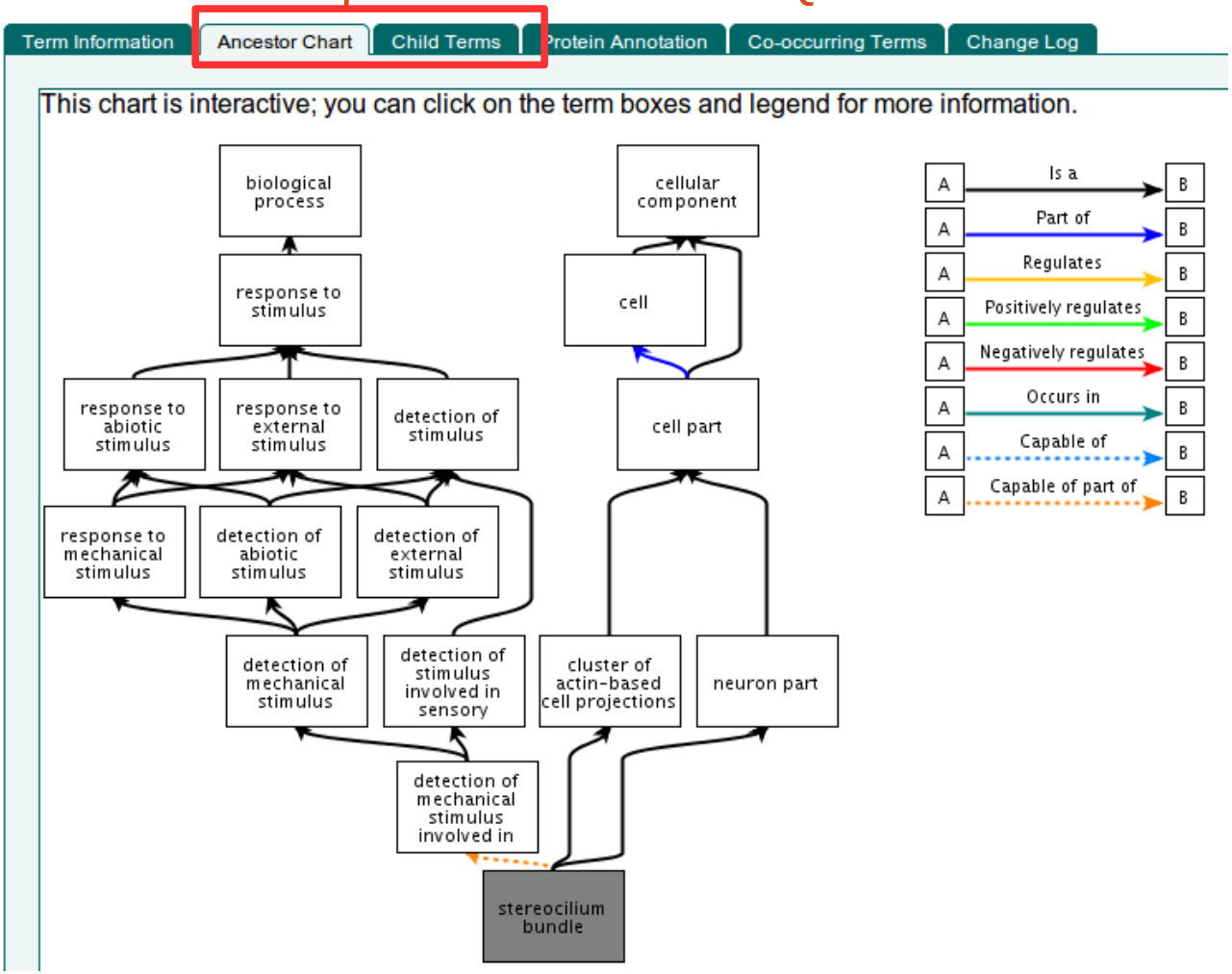
- Import the Gene Ontology
  - Download the basic Gene Ontology
    http://purl.obolibrary.org/obo/go/go-basic.obo
  - Parse the file and find all terms

```
[Term]
id: GO:0032421
name: stereocilium bundle
namespace: cellular_component
def: "A bundle of cross-linked (...)" [GOC:ecd, PMID:15661519, PMID:7840137]
subset: goslim_pir
synonym: "stereocilia bundle" EXACT []
is_a: GO:0097458 ! neuron part
is_a: GO:0098862 ! cluster of actin-based cell projections
```

  - Disregard all terms with „is_obsolete: true“
  - „Connect“ the data to the Gene2Go data from assignment 2 via ID

Task 2: Integrate the Gene Ontology (3)

Helpful resource: https://www.ebi.ac.uk/QuickGO/



Task 3: Queries

- Formulate and execute the following queries:
  - **Task 1**
    (1) Number of distinct human genes after deduplication

    (2) Maximal number of different positions (start and end) assigned to a gene

    (3) How many genes have only one position (start and end)?
  - **Task 2**
    (4) How many GO terms are there in total and for each ontology?

    (5) Compute a frequency histogram over the number of terms assigned to a gene (how many have 0 terms, 1 term, 2 terms, …)

    (6) Write a program (or query) which computes for each term to how many genes it is assigned
    - Store the result in a table ASSIGNMENTS(term_id, count)
    - This is not trivial – you need to traverse a large DAG
    - Use JAVA/Python program, Postgres, recursive queries, …
    - Traverse the tree or use some clever materialization

Competition

- Compute the last query as fast as possible
  (i.e. fill the table ASSIGNMENT as fast as possible)

- To take part, you need to send me your program
  - JAVA/Python program (must be executable „as is“ on gruenau2)
    (or)
  - Postgres query or the name of a PL/pgSQL procedure in your schema

- Time measurement: time (sys+user) at Unix or
  the prcedure execution time in Postgres

Submit

- By Monday, 06.06.2016, 23:59 pm

- Send by mail to: yvonne.lichtblau@informatik.hu-berlin.de (questions are also welcome!)

- **Submit:**
  - An updated schema graph with the new table(s) for GO
  - Queries 1-5: query + result in a text file/PDF file
  - Query 6: Report the value for „biological_process“ (GO:0008150) and leave the full table ASSIGNMENT in your schema
  - Notes of your workflow as PDF

- **Criteria for passing the exercise:**
  Submit everything from the list and correct answer of each query.