BINF 8211/6211
Design and Implementation of Bioinformatics Databases
Lecture #15

Dr. D. Andrew Carr
Dept. Bioinformatics and Genomics UNCC
Spring 2016

Warm-up questions

- What are the Normal Forms in a database design context?
- How do the Normal Forms relate to the mathematical definition of determination?
- Explain each normal form with an example you might see in a spread sheet.
- What are partial dependencies in a table and how to they relate to the Normal Forms?
- What are transitive dependencies in a table and how do they related to the Normal Forms?
- What is an easy way to be sure you are in 2NF?
- What is an easy way to be sure that BCNF = 3NF?
- What is the inevitable result of this process?

Today Schedule

- Homework 5 Assigned
- Discussion of Rest of Course.
- SQL Concepts
- Introduction to Ontologies

Rest of Course

- Homework 5 Due March 29th
- Homework 6 Due April 5th
- Homework 7 Due April 7th

- Presentations Start on the 14th of April
- Final Write ups are due on the 28th of April

The GROUP BY + HAVING clauses let you *filter* data: one thing you can use this for is to generate frequency distributions from the SELECT statements *taken over aggregate functions*.

- SELECT columnlist
- FROM tablelist
- [WHERE conditionlist]
- [GROUP BY columnlist]
- [HAVING conditionlist]
- [ORDER BY columnlist[ASC|DESC]];

GENEOPERONLENGTH

operon_id	gene_id	gene_length
2	deoA	1322
2	deoB	1223
1	moaA	989
2	deoC	779
2	deoD	719
3	flgA	659
1	moaB	512
1	moaC	485
1	moaE	452
3	flgN	416
3	flgM	293
1	moaD	345

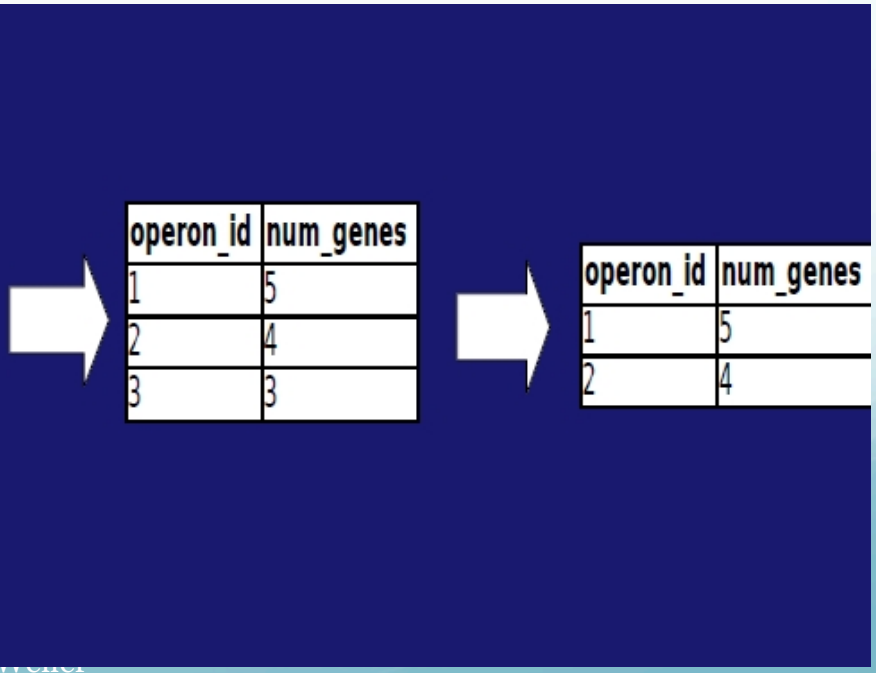
```
SELECT operon_id, COUNT(DISTINCT (gene_id))
AS num_genes
FROM GENEOPERONLENGTH
GROUP BY operon_id
ORDER By num_genes DESC;

operon_id    num_genes
1            5
2            4
3            3
```

Logical query: how many of the genes in the gene table belong to operons with more than 3 genes?

```
SELECT      operon_id,
            COUNT(operon_id) as num_genes
FROM        gene
GROUP BY    operon_id
HAVING      num_genes > 3
ORDER BY    operon_id ASC
```

gene_symbol (PK)	genome_start_loc	genome_end_loc	genome_strand	gene_seq	operon_id (FK)
moaA	816267	817256	+	ATGGCTTCAC	1
moaB	817278	817790	+	ATGAGTCAGG	1
moaC	817793	818278	+	ATGTCGCAAC	1
moaD	818271	818516	+	ATGATTAAAG	1
moaE	818518	818970	+	ATGGCAGAAA	1
flgN	1128637	1129053	-	ATGACACGTC	3
flgM	1129058	1129351	-	ATGAGTATTG	3
flgA	1129427	1130086	-	ATGCTGATAA	3
deoC	4615346	4616125	+	ATGACTGATC	2
deoA	4616252	4617574	+	TTGTTTCTCG	2
deoB	4617626	4618849	+	ATGAAACGTG	2
deoD	4618906	4619625	+	ATGGCTACCC	2



operon_id	num_genes
1	5
2	4
3	3

operon_id	num_genes
1	5
2	4

SQL

- Subquery
  - In the WHERE or HAVING clause
- Vs.
- Inline Query
  - In the FROM

SQL concepts – Subquery

A subquery is a SELECT statement in the WHERE or HAVING clause of *another* SELECT statement. The subquery executes *first* and feeds output into the main query.

```
SELECT ename, deptno
   FROM emp
    WHERE deptno = (SELECT deptno
                                  FROM emp
                                  WHERE ename = 'TAYLOR');
```

```
SELECT ename, deptno, sal
   FROM emp, x
   WHERE sal > (SELECT AVG(sal)
                        FROM emp
                        WHERE emp.deptno = x.deptno)
ORDER BY deptno;
```

SQL concepts – The inline view is a subset of subqueries that use the FROM part of the statement.

The *inline view*, or *derived* table is a subquery (SELECT) inside another query that placed inside the FROM clause to be used as a *run-time result set*.

The view only exists in the query in which it is created.

It can be used to simplify complex queries by removing JOIN operations

```
SELECT *
   FROM ( SELECT deptno, count(*) emp_count
               FROM emp
                   GROUP BY deptno ) emp, dept
WHERE dept.deptno = emp.deptno;
```

```
SELECT a.last_name, a.salary, a.department_id, b.maxsal
    FROM employees a,
            ( SELECT department_id, max(salary) maxsal
                  FROM employees
                  GROUP BY department_id ) b
WHERE a.department_id = b.department_id AND a.salary = b.maxsal;
```

GENEOPERONLENGTH

operon_id	gene_id	gene_length
2	deoA	1322
2	deoB	1223
1	moaA	989
2	deoC	779
2	deoD	719
3	flgA	659
1	moaB	512
1	moaC	485
1	moaE	452
3	flgN	416
3	flgM	293
1	moaD	345

```
SELECT operon_id, gene_id, gene_length
FROM GENEOPERONLENGTH
WHERE gene_length>(SELECT AVG (gene_length) FROM GENEOPERONLENGTH);

operon_id    gene_id      gene_length
2        deoA     1322
2        deoB     1233
1        maoA     989
2        deoC     779
2        deoD     719
```

Ontology

- Philosophical study of:
  - Being
  - Becoming
  - Existence or reality
  - Categorization or being and their relations.
- Metaphysics.
  - What entities exist or may be said to exist.
  - Groupings, heirarchy
- Practical application in developing relationships

Ontology

- Principal questions of ontology include:
- "What can be said to exist?"
- "What is a thing?”
- "Into what categories, if any, can we sort existing things?"
- "What are the meanings of being?"
- "What are the various modes of being of entities?"

Ontology

- Representations from natural language are squishy

- Natural languages are expressive - the symbols and rules for manipulating them have a range, but this is often not defined. The result is a lot of ambiguity.
  - Humans have low limits on how much information can be processed and manipulated.
- Computing requires explicit rules and boundaries, but can process/manipulate far more information.
- The language in which symbols and rules are expressed is a limit for both.

Why ontologies?

- Keeping things straight....

- What is a biological sequence.
  - What is the importance of the sequence?

Ontologies: formalizations of a language domain, representing *knowledge*

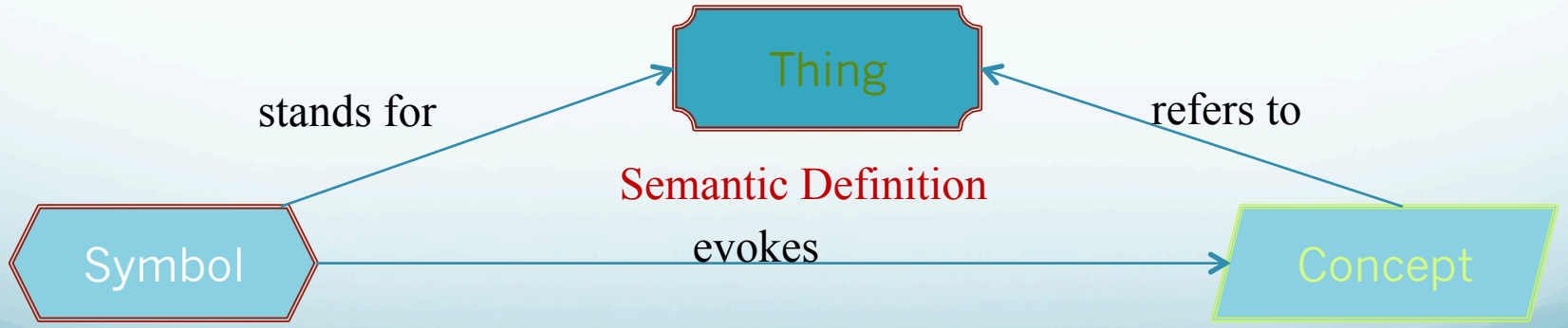
- Biological and Biomedical Ontologies – introduction and examples
- The OBO and OWL ontology frameworks
- Reading: Chapter 4 in Introduction to Bioontologies by Robinson and Bauer
- The Bio-Ontologies SIG is a discussion forum (last year seems to be 2013 however) – possibly most useful is the list of conference papers http://bio-ontologies.knowledgeblog.org/table-of-contents . I like the Processes and Properties paper. In 2015 the SIG seems to be here http://www.bio-ontologies.org.uk/ but I don't see proceedings posted yet.
- The BioPortal collects and manages biomedical ontologies http://bioportal.bioontology.org/
- An finally, Barry Smith is a philosopher who helped get the application of ontologies in the biological research domain started and he has a large number of helpful papers explaining pretty much any aspect of the process. I note here on 'How to Distinguish Parthood fromLocation in Bio-Onotlogies" http://ontology.buffalo.edu/bio/Part&Location.pdf .

Bio-ontologies focus on problems of biological data representation: naming the terms (*annotation)* and relationships *(rules and policies allowing integration)* in the creation of knowledge.

- Goals: data exchange and data mining
  - Data exchange
    - Sharing annotations but also *predicting* correct annotations
    - Predicting new relationships as well as paths between nodes
    - Grouping data by shared annotations
    - Identifying key words that allow productive text mining
  - Data mining
    - Identify correlations
    - Clustering data based on similar properties

Things, concepts, symbols

- A *real thing* has mass and other attributes.
- Our *concept* of that thing usually simplifies (we identify key attributes and ignore the others) and also specifies/ generalizes.
- We build concepts using *symbols*, including symbols for the attributes and their allowed values.



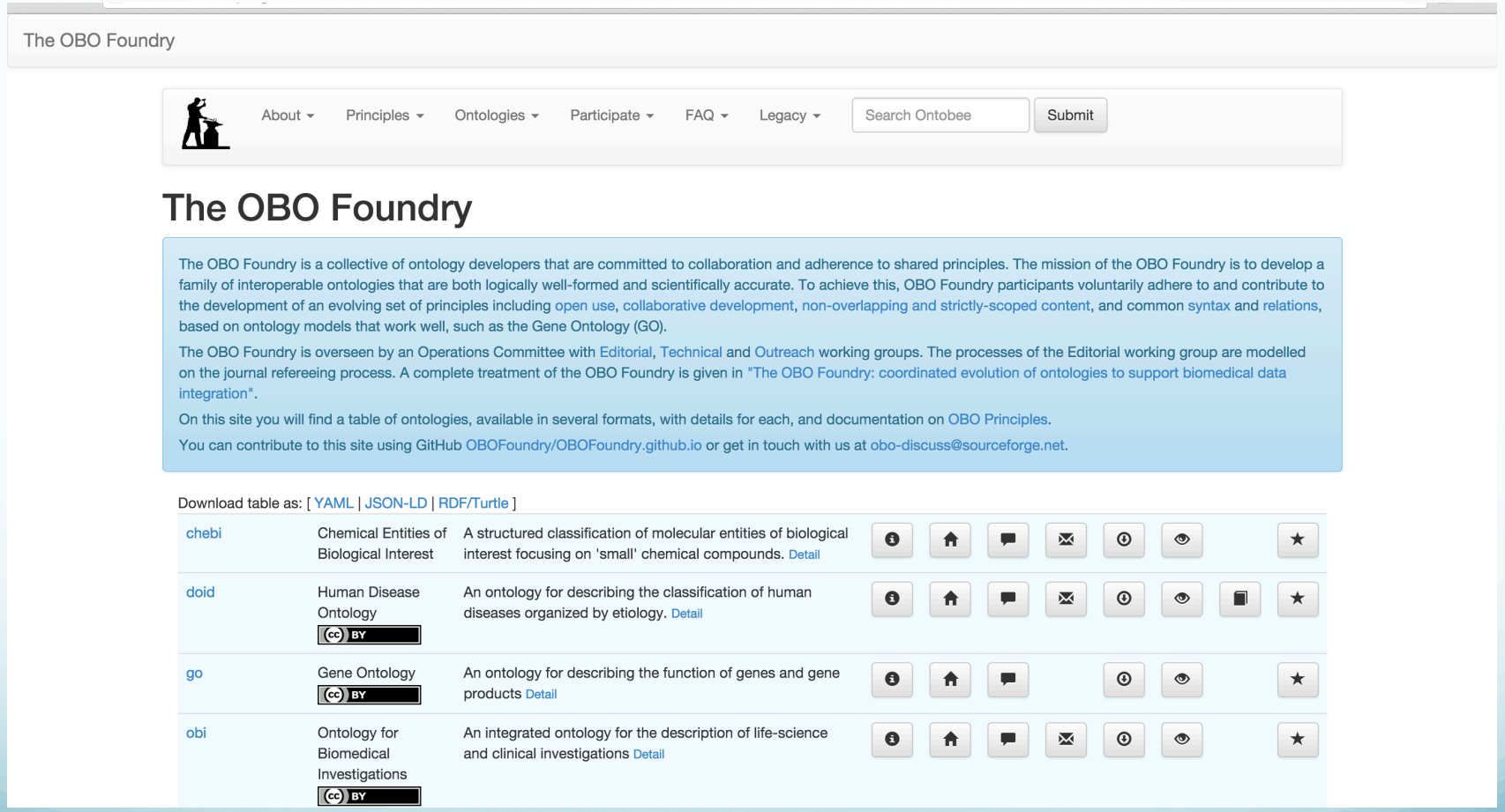
Ontologies

- In the sciences we create formal representations of the concepts in a particular domain
  - We represent (use symbols for) the terms for things, their properties, and the allowed relationships between things.
  - Accuracy and precision are mutual goals
- Concepts are the ‘units’, terms are the labels, and relationships are the links between units.
  - Units are classes and individuals
  - Frames are the properties
  - Relationships are predicates and describe facts about the classes and their properties.

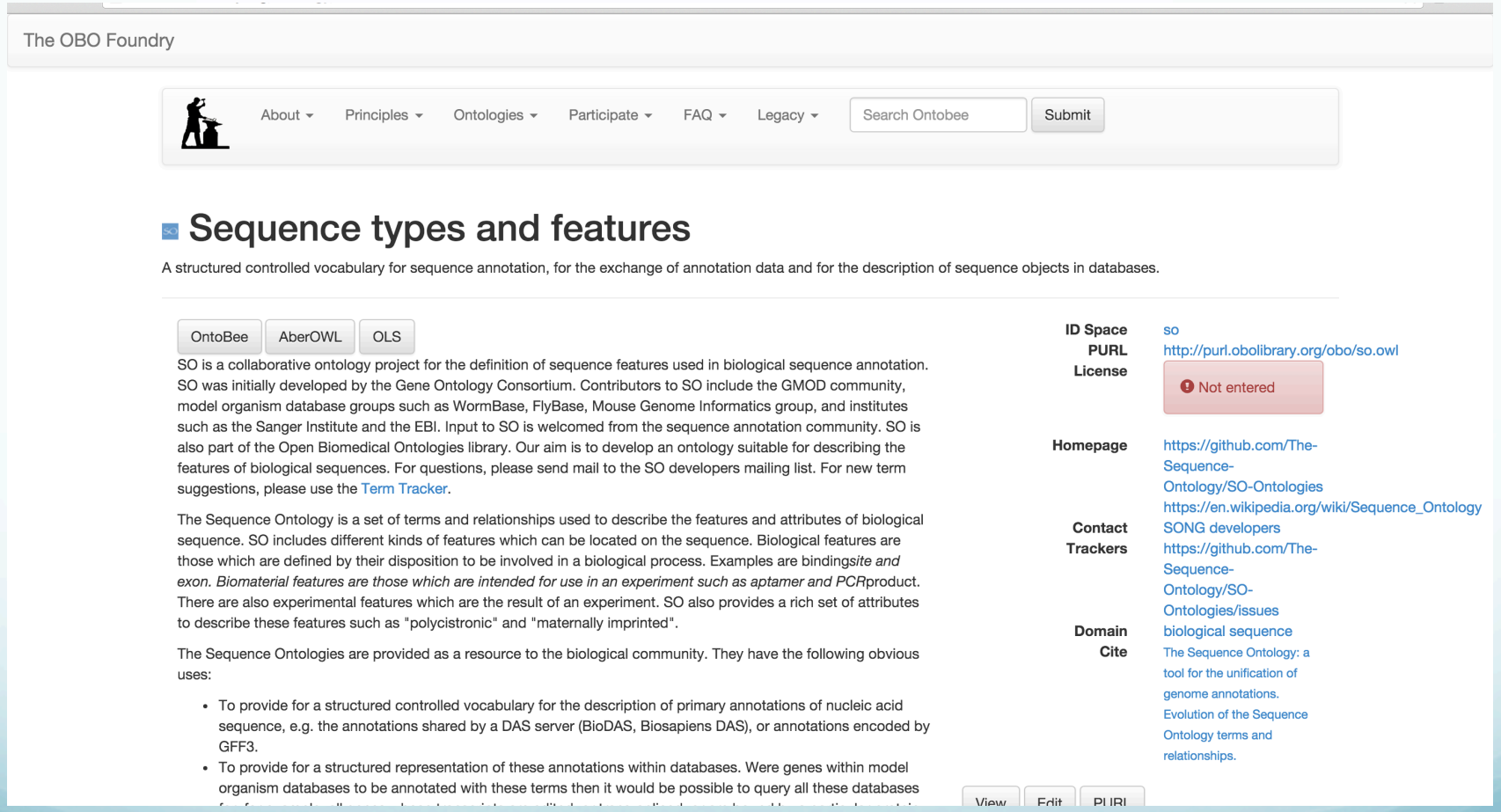
Topics (extended)

- Ontologies
- Bio-ontologies
- The Gene Ontology
- The Sequence Ontology
  - The Genbank Flatfile 3 (GFF3) format

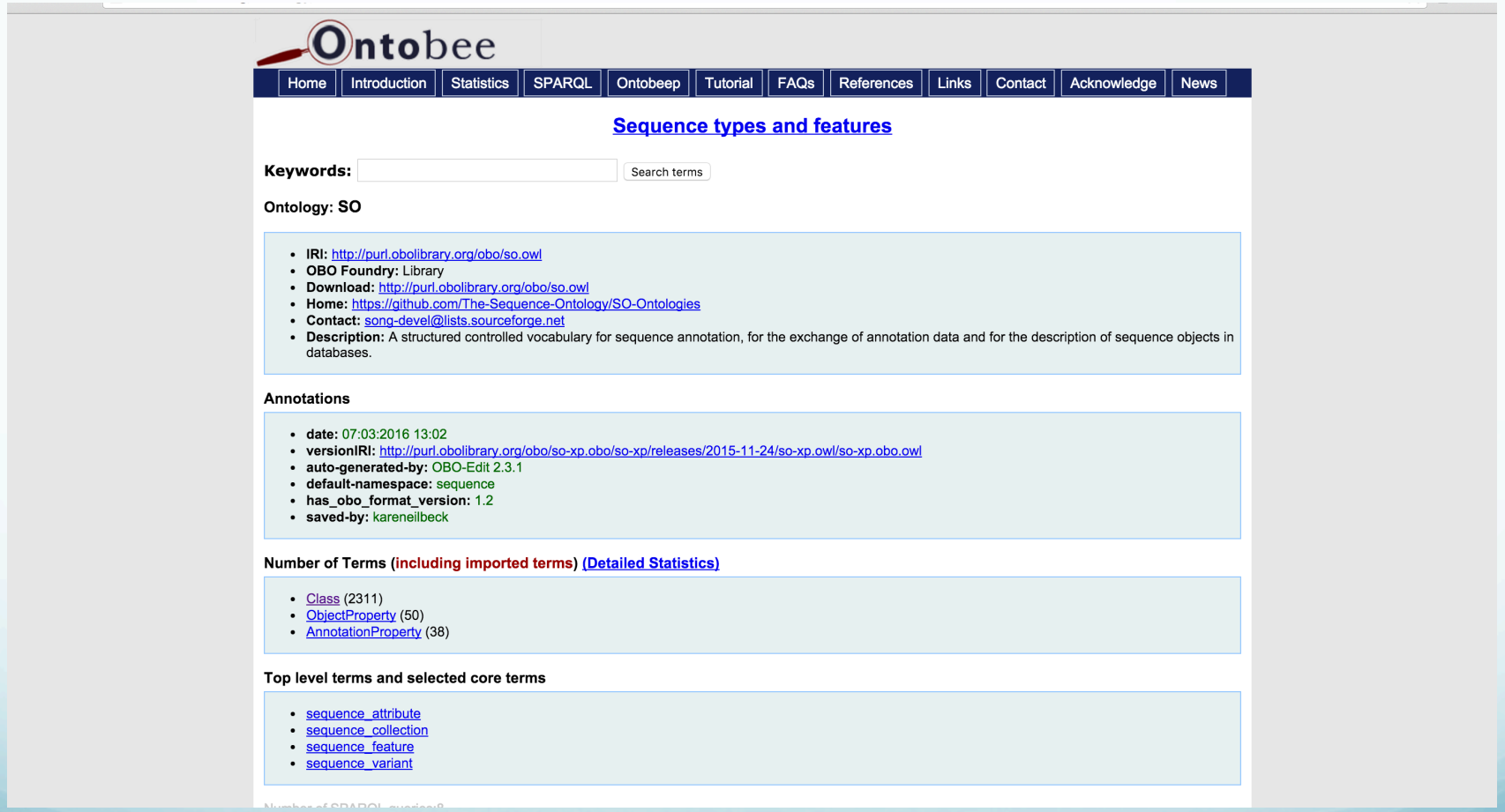
OBO Foundary



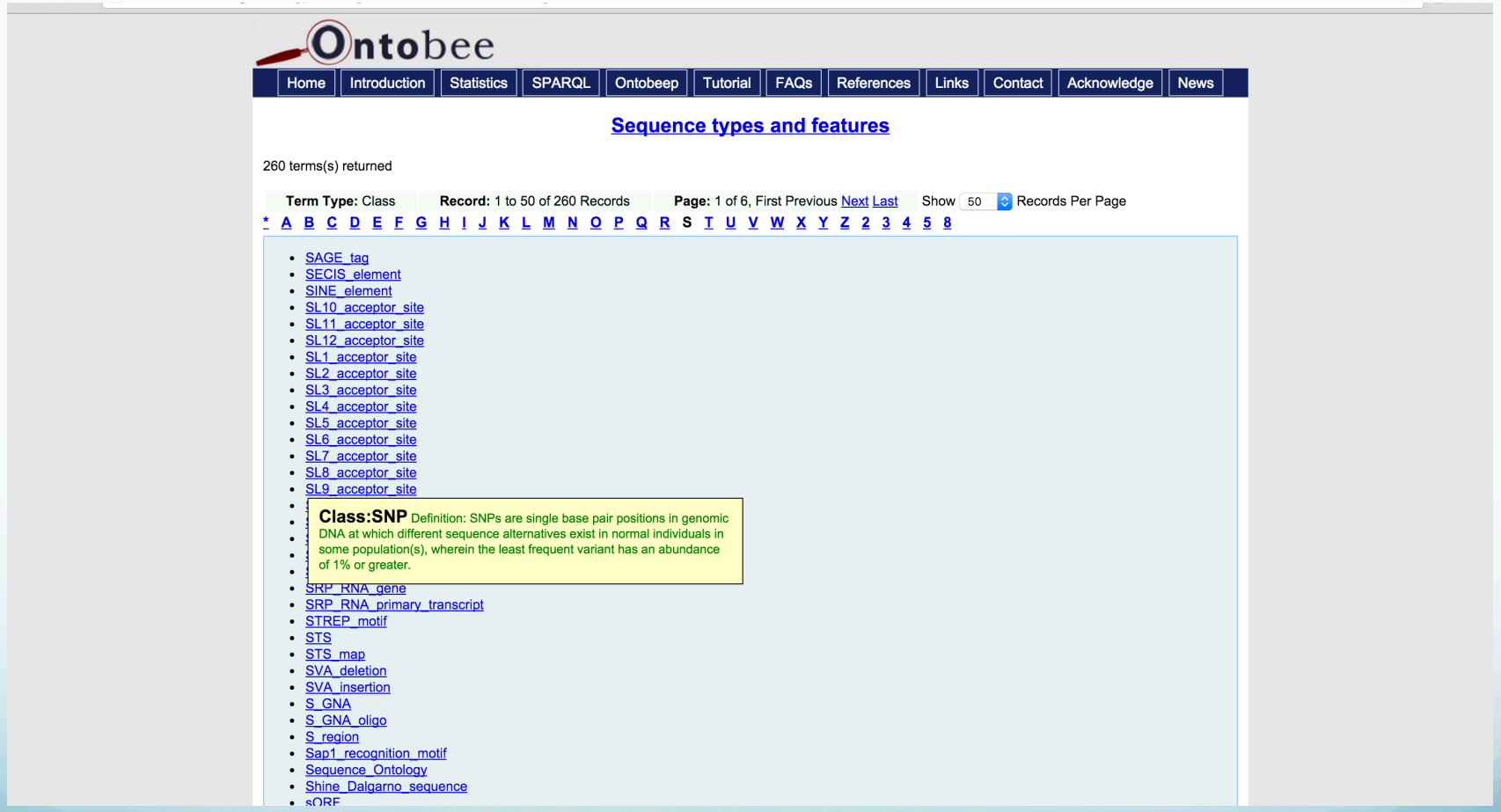
OBO Foundary SO



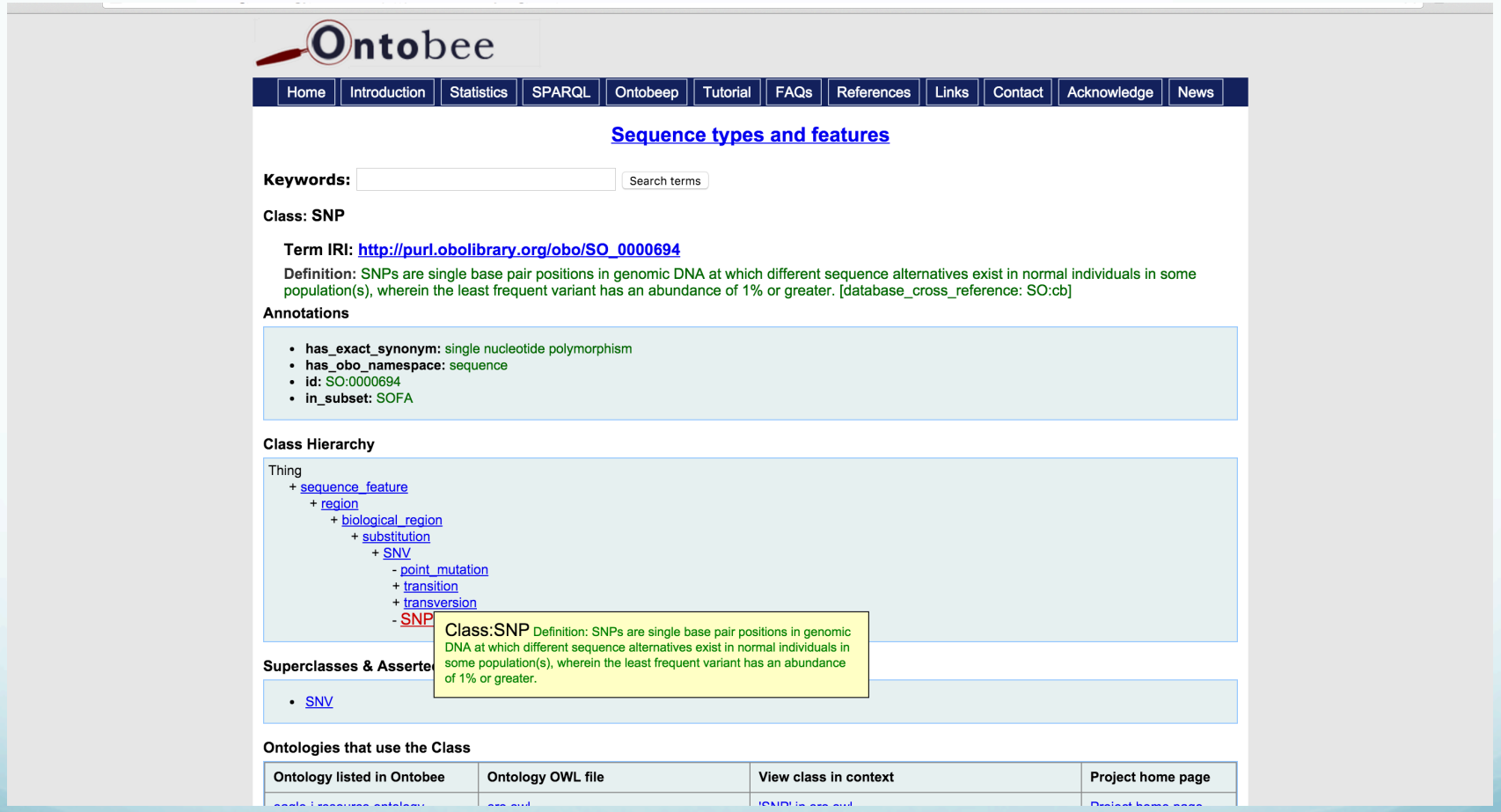
OBO OntoBee

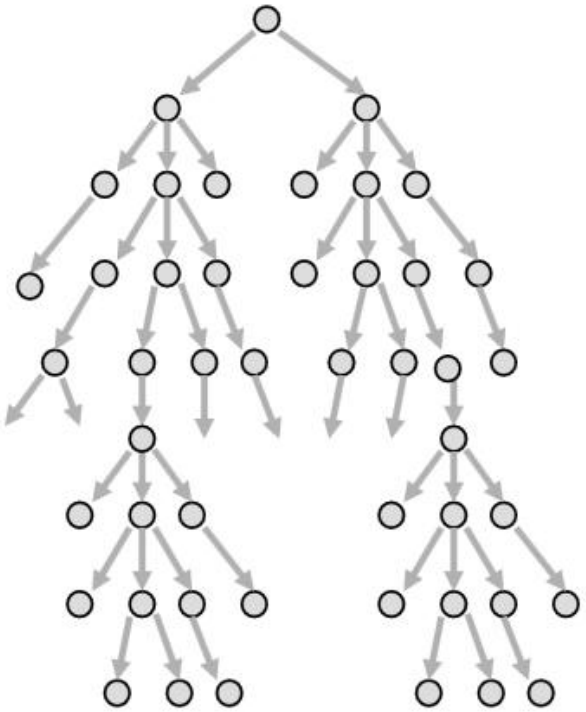

Ontobee
Home
Introduction
Statistics
SPARQL
Ontobeep
Tutorial
FAQs
References
Links
Contact
Acknowledge
News
Sequence types and features
Keywords:
Search terms
Ontology: SO
IRI: http://purl.obolibrary.org/obo/so.owl
OBO Foundry: Library
Download: http://purl.obolibrary.org/obo/so.owl
Home: https://github.com/The-Sequence-Ontology/SO-Ontologies
Contact: song-devel@lists.sourceforge.net
Description: A structured controlled vocabulary for sequence annotation, for the exchange of annotation data and for the description of sequence objects in databases.
Annotations
date: 07:03:2016 13:02
versionIRI: http://purl.obolibrary.org/obo/so-xp.obo/so-xp/releases/2015-11-24/so-xp.owl/so-xp.obo.owl
auto-generated-by: OBO-Edit 2.3.1
default-namespace: sequence
has_obo_format_version: 1.2
saved-by: kareneilbeck
Number of Terms (including imported terms) (Detailed Statistics)
Class (2311)
ObjectProperty (50)
AnnotationProperty (38)
Top level terms and selected core terms
sequence_attribute
sequence_collection
sequence_feature
sequence_variant

OBO SO ONTOBEE SNP



OBO SO SNP continued

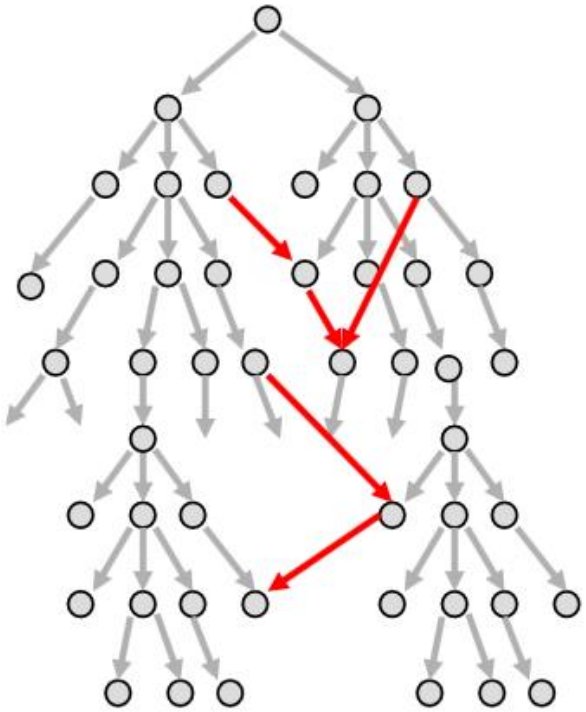




Rule "has part"

Directed rule: 1 parent

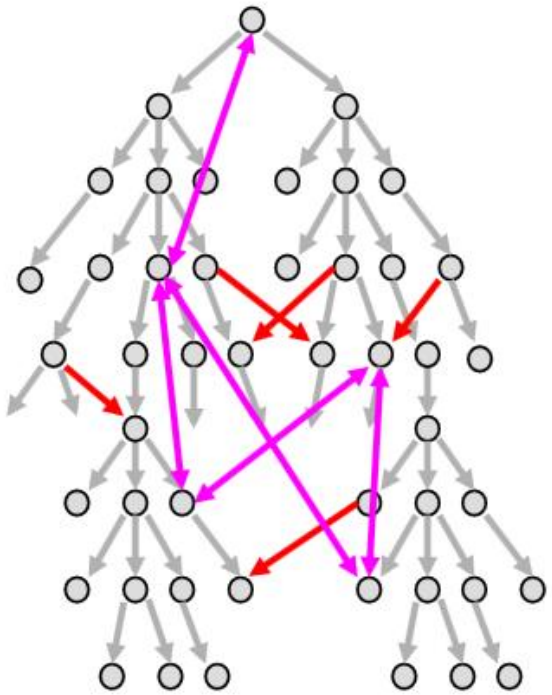
Simple hierarchy



Rule "signals to" with wnt protein

Directed rule: >1 parent

Directed acyclic graph



Rule "is next to"

Undirected rule and parents ≡ children

Graph