

BINF 8211/6211

Design and Implementation of Bioinformatics Databases

Lecture #15

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

- 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?

Warm-up questions

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
figN	1128637	1129053	-	ATGACACGTC	3
figM	1129058	1129351	-	ATGAGTATTG	3
figA	1129427	1130086	-	ATGCTGATAA	3
deoC	4615346	4616125	+	ATGACTGATC	2
deoA	4616252	4617574	+	TTGTTTCTCG	2
deoB	4617626	4618849	+	ATGAAACGTG	2
deoD	4618906	4619625	+	ATGGCTACCC	2

The diagram illustrates the logical flow of data from the gene table to the final result. It starts with the gene table on the left, which is processed by a SELECT statement to produce an intermediate table in the middle. This intermediate table is then processed by another SELECT statement to produce the final result table on the right.

operon_id	num_genes
1	5
2	4
3	3

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	1223
1	moaA	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, hierarchy
- 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*

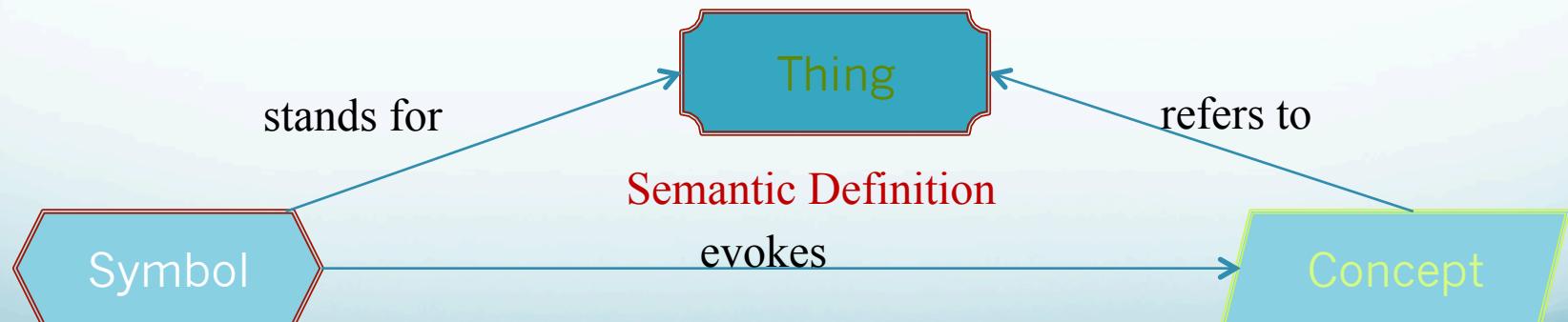
- Biological and Biomedical Ontologies – introduction and examples
- The OBO and OWL ontology frameworks
- Reading: Chapter 4 in Introduction to Bioontologies by Robinson and Bauer
- The Bio-Ontologies SIG is a discussion forum (last year seems to be 2013 however) – possibly most useful is the list of 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://biportal.bioontology.org/>
- And 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 from Location in Bio-Ontologies’
<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.

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

Topics (extended)

OBO Foundry

The OBO Foundry



About ▾

Principles ▾

Ontologies ▾

Participate ▾

FAQ ▾

Legacy ▾

Search Ontobee

Submit

The OBO Foundry

The OBO Foundry is a collective of ontology developers that are committed to collaboration and adherence to shared principles. The mission of the OBO Foundry is to develop a family of interoperable ontologies that are both logically well-formed and scientifically accurate. To achieve this, OBO Foundry participants voluntarily adhere to and contribute to the development of an evolving set of principles including open use, collaborative development, non-overlapping and strictly-scoped content, and common syntax and relations, based on ontology models that work well, such as the Gene Ontology (GO).

The OBO Foundry is overseen by an Operations Committee with Editorial, Technical and Outreach working groups. The processes of the Editorial working group are modelled on the journal refereeing process. A complete treatment of the OBO Foundry is given in "The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration".

On this site you will find a table of ontologies, available in several formats, with details for each, and documentation on OBO Principles.

You can contribute to this site using GitHub [OBOFoundry/OBOFoundry.github.io](#) or get in touch with us at obo-discuss@sourceforge.net.

Download table as: [[YAML](#) | [JSON-LD](#) | [RDF/Turtle](#)]

chebi	Chemical Entities of Biological Interest	A structured classification of molecular entities of biological interest focusing on 'small' chemical compounds. Detail	
doid	Human Disease Ontology	An ontology for describing the classification of human diseases organized by etiology. Detail	
go	Gene Ontology	An ontology for describing the function of genes and gene products Detail	
obi	Ontology for Biomedical Investigations	An integrated ontology for the description of life-science and clinical investigations Detail	

OBO Foundry SO

The OBO Foundry



About ▾

Principles ▾

Ontologies ▾

Participate ▾

FAQ ▾

Legacy ▾

Search Ontobee

Submit

Sequence types and features

A structured controlled vocabulary for sequence annotation, for the exchange of annotation data and for the description of sequence objects in databases.

OntoBee AberOWL OLS

SO is a collaborative ontology project for the definition of sequence features used in biological sequence annotation. SO was initially developed by the Gene Ontology Consortium. Contributors to SO include the GMOD community, model organism database groups such as WormBase, FlyBase, Mouse Genome Informatics group, and institutes such as the Sanger Institute and the EBI. Input to SO is welcomed from the sequence annotation community. SO is also part of the Open Biomedical Ontologies library. Our aim is to develop an ontology suitable for describing the features of biological sequences. For questions, please send mail to the SO developers mailing list. For new term suggestions, please use the [Term Tracker](#).

The Sequence Ontology is a set of terms and relationships used to describe the features and attributes of biological sequence. SO includes different kinds of features which can be located on the sequence. Biological features are those which are defined by their disposition to be involved in a biological process. Examples are *bindingsite* and *exon*. *Biomaterial features* are those which are intended for use in an experiment such as *aptamer* and *PCRproduct*. There are also experimental features which are the result of an experiment. SO also provides a rich set of attributes to describe these features such as "polycistronic" and "maternally imprinted".

The Sequence Ontologies are provided as a resource to the biological community. They have the following obvious uses:

- To provide for a structured controlled vocabulary for the description of primary annotations of nucleic acid sequence, e.g. the annotations shared by a DAS server (BioDAS, Biosapiens DAS), or annotations encoded by GFF3.
- To provide for a structured representation of these annotations within databases. Were genes within model organism databases to be annotated with these terms then it would be possible to query all these databases

ID Space

PURL

License

so

<http://purl.obolibrary.org/obo/so.owl>

ⓘ Not entered

Homepage

<https://github.com/The-Sequence-Ontology/SO-Ontologies>

https://en.wikipedia.org/wiki/Sequence_Ontology

Contact

Trackers

<https://github.com/The-Sequence-Ontology/SONG-developers>

<https://github.com/The-Sequence-Ontology/SO-Ontologies/issues>

Domain

Cite

<https://biologicalsequence.org/>

[The Sequence Ontology: a tool for the unification of genome annotations](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8030333/)

[Evolution of the Sequence Ontology terms and relationships](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8030333/#sec1)

View

Edit

PI IRI

OBO OntoBee

The Ontobee logo features the word "Ontobee" in a stylized font where the "O" is replaced by a red swoosh.

[Home](#) | [Introduction](#) | [Statistics](#) | [SPARQL](#) | [OntobeeP](#) | [Tutorial](#) | [FAQs](#) | [References](#) | [Links](#) | [Contact](#) | [Acknowledge](#) | [News](#)

Sequence types and features

Keywords:

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: karenellbeck

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](#)

Number of SPARQL queries: 8

OBO SO ONTOBEE SNP

 Ontobee

Home | Introduction | Statistics | SPARQL | OntoBEEP | Tutorial | FAQs | References | Links | Contact | Acknowledge | News

Sequence types and features

260 terms(s) returned

Term Type: Class Record: 1 to 50 of 260 Records Page: 1 of 6, First Previous [Next](#) [Last](#) Show Records Per Page

* A B C D E F G H I J K L M N O P Q R S T U V W X Y Z 2 3 4 5 8

- [SAGE_tag](#)
- [SECIS_element](#)
- [SINE_element](#)
- [SL10_acceptor_site](#)
- [SL11_acceptor_site](#)
- [SL12_acceptor_site](#)
- [SL1_acceptor_site](#)
- [SL2_acceptor_site](#)
- [SL3_acceptor_site](#)
- [SL4_acceptor_site](#)
- [SL5_acceptor_site](#)
- [SL6_acceptor_site](#)
- [SL7_acceptor_site](#)
- [SL8_acceptor_site](#)
- [SL9_acceptor_site](#)
- **Class:SNP** Definition: SNPs are single base pair positions in genomic DNA at which different sequence alternatives exist in normal individuals in some population(s), wherein the least frequent variant has an abundance of 1% or greater.
- [SRP_RNA_gene](#)
- [SRP_RNA_primary_transcript](#)
- [STREP_motif](#)
- [STS](#)
- [STS_map](#)
- [SVA_deletion](#)
- [SVA_insertion](#)
- [S_GNA](#)
- [S_GNA_oligo](#)
- [S_region](#)
- [Sap1_recognition_motif](#)
- [Sequence_Ontology](#)
- [Shine_Dalgarno_sequence](#)
- [sORF](#)

OBO SO SNP continued

 [Ontobee](#)

[Home](#) [Introduction](#) [Statistics](#) [SPARQL](#) [OntoBeep](#) [Tutorial](#) [FAQs](#) [References](#) [Links](#) [Contact](#) [Acknowledge](#) [News](#)

[Sequence types and features](#)

Keywords: [Search terms](#)

Class: SNP

Term IRI: http://purl.obolibrary.org/obo/SO_0000694

Definition: SNPs are single base pair positions in genomic DNA at which different sequence alternatives exist in normal individuals in some population(s), wherein the least frequent variant has an abundance of 1% or greater. [database_cross_reference: SO:cb]

Annotations

- **has_exact_synonym:** single nucleotide polymorphism
- **has_obo_namespace:** sequence
- **id:** SO:0000694
- **in_subset:** SOFA

Class Hierarchy

```
graph TD; Thing --> sequence_feature[sequence_feature]; sequence_feature --> region[region]; region --> biological_region[biological_region]; biological_region --> substitution[substitution]; substitution --> SNV[SNV]; SNV --> point_mutation[point_mutation]; SNV --> transition[transition]; SNV --> transversion[transversion]; SNV --> SNP[SNP]
```

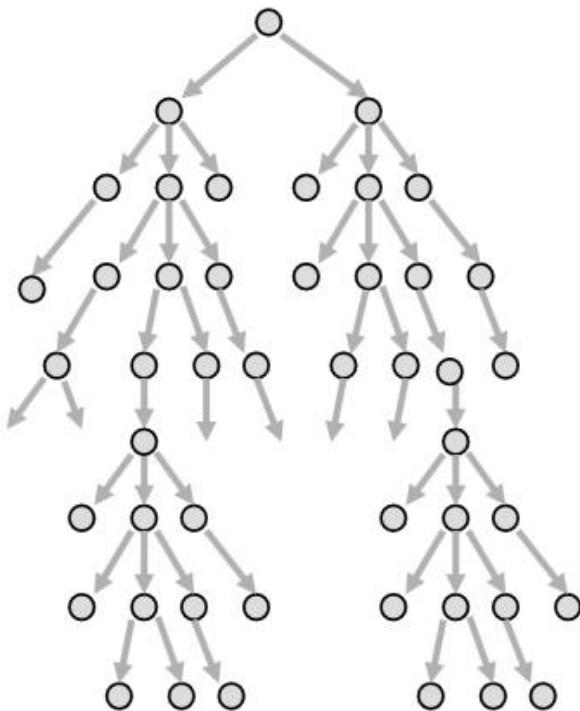
Class:SNP Definition: SNPs are single base pair positions in genomic DNA at which different sequence alternatives exist in normal individuals in some population(s), wherein the least frequent variant has an abundance of 1% or greater.

Superclasses & Assertions

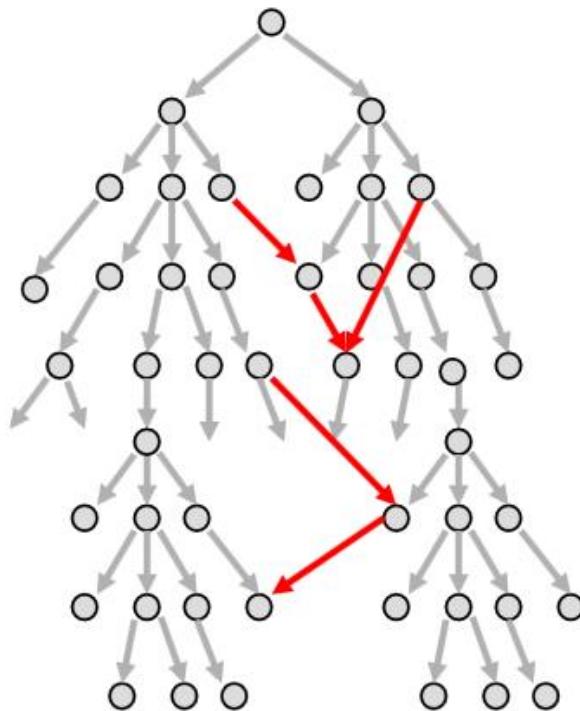
- [SNV](#)

Ontologies that use the Class

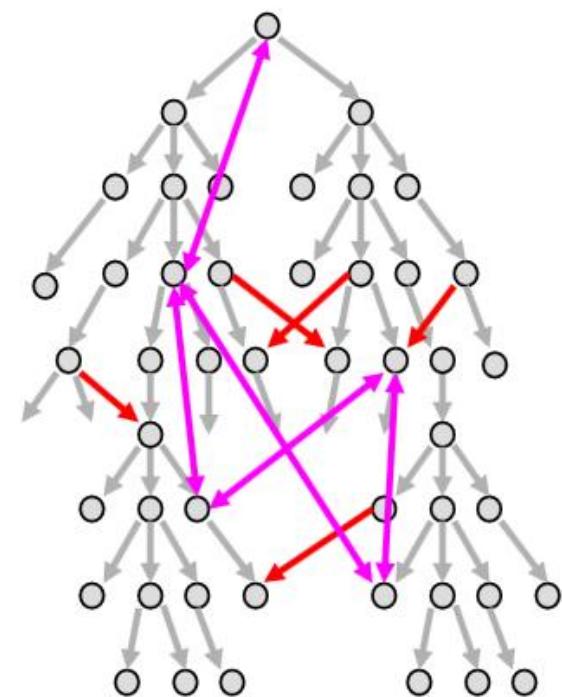
Ontology listed in Ontobee	Ontology OWL file	View class in context	Project home page
single nucleotide polymorphism	owl.owl	SO.owl	Project home page



→ 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