How ontologies are used in life sciences

Challenges in biology and the role of ontologies

James Malone
malone@ebi.ac.uk
Outline

• Challenges in biology and things we describe
• Use cases for ontologies in biology
• Examples of how we use them in biology
Biology represents a diverse set of topics:

- Genomes
- Nucleotide sequence
- Gene expression
- Protein families, domains and motifs
- Protein-protein interactions
- Pathways
- Literature
- Protein sequence
- Proteomes
- Protein structure
- Chemical entities
- Systems
....each of which produces data

Genomes
- Ensembl
- Ensembl Genomes
- EGA

Nucleotide sequence
- ENA

Functional genomics
- ArrayExpress
- Expression Atlas

Protein activity
- IntAct
- PRIDE

Protein Sequences
- UniProt

Chemical entities
- ChEBI

Chemogenomics
- ChEMBL

Literature and ontologies
- CiteXplore
- GO

Protein families, motifs and domains
- InterPro

Macromolecular
- PDBe

Pathways
- Reactome

Systems
- BioModels
- BioSamples
Would be nice to put all this stuff together…
… but it’s hard
Many existing challenges in life sciences

- Understanding connections between people and diseases
  - gene-disease, population-disease
- Identifying new drug targets
  - Faster, cheaper with less side effects
- Tailoring treatment to the individual
  - Better outcomes, lower adverse events
- Understanding relationship between human and animal
  - Genetic homology
- Predicting protein function
Problems Ontologies in Biology Try To Solve

• Provenance – where did it come from, who did it?
• Reproducibility – can I repeat and find results reported?
• Sharing – can others understand your data?
• Integration – can I readily take multiple (thousands of) data sets and use them without preparation?
• New knowledge – can we infer new knowledge as a sum of current knowledge (computationally)?
Provenance

- Ontologies provide a structured, controlled mechanism to capture provenance - who said it? When? How did they reach this conclusion?
- **OBO Evidence Code Ontology** (ECO) – for making statements about scientific evidence for conclusion, e.g. ‘match to sequence model’
- **W3C Provenance Ontology** (PROV-O) – set of general classes and relations for capturing provenance, e.g. ‘wasAttributedTo’ (some) ‘Person’
- **W3C Open Annotation Model** (OA) – for sharing annotations, e.g. Annotation hasBody ‘liver cancer’
Reproducibility

- Making results reproducible is to fully describe methods
- OBI attempts to help with this
- Several papers have discusses problems when experiments replicated
Reproducibility – Software Registries

- BMB software registry which uses SWO-EDAM
Reproducibility – Workflows in Taverna

- myGrid has ontologies for semantic service discovery (the myGrid ontology) and for managing process provenance (the workflow run ontology)
- I know less about this – speak to Robert though
Sharing - Databases using ontologies

- Mouse Genome Informatics (MGI) and Edinburgh Mouse Atlas (EMAP) (see Terry’s talk tomorrow)
  - GO, Mammalian Phenotype, Mouse Gross Anatomy
- ArrayExpress and Gene Expression Atlas
  - EFO which uses GO, ChEBI, Cell Type, PATO, UO, OBI and more
- ISATab framework
  - Ontology agnostic
Sharing – Edinburgh Mouse Atlas Project
http://www.emouseatlas.org
Sharing – Virtual Fly Brain
www.virtualflybrain.org
Sharing - Databases using ontologies – UniProt

www.uniprot.org/
### Gene Ontology (GO)

#### Biological process

- **DNA damage response, signal transduction by p53 class mediator resulting in transcription of p21 class mediator**  
  Traceable author statement (PubMed 10918303), Source: UniProtKB

- **G2/M transition DNA damage checkpoint**  
  Inferred from mutant phenotype (Ref.41 Ref.39 Ref.45), Source: UniProtKB

- **Androgen receptor signaling pathway**  
  Non-traceable author statement (PubMed 15572661), Source: UniProtKB

- **Cell cycle**  
  Inferred from electronic annotation. Source: UniProtKB-KW

- **Cellular response to indole-3-methanol**  
  Inferred from direct assay (PubMed 10886478), Source: UniProtKB

- **Chromosome segregation**  
  Inferred from mutant phenotype (PubMed 15995487), Source: UniProtKB

- **Double-strand break repair via homologous recombination**  
  Inferred from direct assay (PubMed 17349954), Source: HGNC

- **Fatty acid biosynthetic process**  
  Inferred from electronic annotation. Source: UniProtKB-KW

- **Intrinsic apoptotic signaling pathway in response to DNA damage**  
  Inferred from direct assay (PubMed 146854789), Source: MGI

- **Negative regulation of centriole replication**  
  Non-traceable author statement (PubMed 12214252), Source: UniProtKB

- **Negative regulation of fatty acid biosynthetic process**  
  Inferred from mutant phenotype (Ref.35), Source: UniProtKB

- **Negative regulation of histone H3-K9 methylation**
Sharing Databases using ontologies - Atlas

**ATLAS**

Find genes matching all of the following conditions:
- in at least ______ exp.
- ______ is up or down
- in ______

Add conditions to the query:
- Gene property
- Condition
- Organism

Genes 1-50 of 25054 total found (you can refine your query) • Download all results • JSON XML

Legend:
- number of studies the gene is over/under expressed in (~ in experiment pop-ups indicates non-differential expression)

<table>
<thead>
<tr>
<th>Gene</th>
<th>Organism</th>
<th>Entrez Gene</th>
<th>ENCODE cell line</th>
</tr>
</thead>
<tbody>
<tr>
<td>RUNX1</td>
<td>Homo sapiens</td>
<td>10</td>
<td>2</td>
</tr>
<tr>
<td>LGALS8</td>
<td>Homo sapiens</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>PPP2R5C</td>
<td>Homo sapiens</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>HTATIP2</td>
<td>Homo sapiens</td>
<td>1</td>
<td>2</td>
</tr>
</tbody>
</table>
Sharing Databases using ontologies - Atlas
Data Integration

- Consider this picture of gene/population – trait SNP associations

Credit: http://www.genome.gov/gwastudies/
GWAS with ontologies
http://www.ebi.ac.uk/fgpt/sw/gwas/

- Query for ‘liver disease’ uses ontology to fetch results
Integration across data – ArrayExpress to Atlas – www.ebi.ac.uk/gxa
Expression Atlas – www.ebi.ac.uk/gxa

Legend: \[ \text{number of studies the gene is over/under expressed in} \] (~ in experiment pop-ups indicates non-differential expression)

<table>
<thead>
<tr>
<th>Gene</th>
<th>Organism</th>
</tr>
</thead>
<tbody>
<tr>
<td>PRMT2</td>
<td>Homo sapiens</td>
</tr>
<tr>
<td>IDS</td>
<td>Homo sapiens</td>
</tr>
<tr>
<td>MLL</td>
<td>Homo sapiens</td>
</tr>
<tr>
<td>SNAP23</td>
<td>Homo sapiens</td>
</tr>
<tr>
<td>YY1</td>
<td>Homo sapiens</td>
</tr>
<tr>
<td>RAP2B</td>
<td>Homo sapiens</td>
</tr>
<tr>
<td>CORO1C</td>
<td>Homo sapiens</td>
</tr>
<tr>
<td>NTN</td>
<td>Homo sapiens</td>
</tr>
</tbody>
</table>

- \[ \text{number of studies the gene is over/under expressed in} \] (~ in experiment pop-ups indicates non-differential expression)
New Knowledge - Analysis using ontologies

• Gene set enrichment analysis – used to determine whether a list of gene is stat sig differences between two states (e.g. disease vs normal)

• Gene ontology enrichment analysis looks for GO annotations shared between these genes

• Also looks for parents using ontology structure

• Offers hints as to what genes may have in common

• Very common analysis step when looking at gene lists
Analysis using ontologies

- Topology and network analysis
- Edge counting between concepts to determine semantic distance
- Interconnections between concepts can be weighted to give interaction scores
**GO supporting functional classification**

**Table 3.** Functional classification of CAN genes, with CaMP score to the right of each gene name. CAN genes were assigned to functional classes using Gene Ontology (GO) groups, INTERPRO domains, and available literature. Representative GO groups and INTERPRO domains are listed for each class.

<table>
<thead>
<tr>
<th>Breast cancers</th>
<th>Colorectal cancers</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Cellular adhesion and motility</strong> (examples: cytoskeletal protein binding GO:0008092, cell adhesion GO:0007155, metallopeptidase activity GO:0008237)</td>
<td></td>
</tr>
<tr>
<td><strong>FLNB</strong></td>
<td>3.4</td>
</tr>
<tr>
<td>MYH1</td>
<td>2.7</td>
</tr>
<tr>
<td>SPTAN1</td>
<td>2.6</td>
</tr>
<tr>
<td>DBN1</td>
<td>2.5</td>
</tr>
<tr>
<td>TECTA</td>
<td>2.4</td>
</tr>
<tr>
<td>ADAM12</td>
<td>2.3</td>
</tr>
<tr>
<td>GSN</td>
<td>2.2</td>
</tr>
<tr>
<td>CDH20</td>
<td>2.2</td>
</tr>
<tr>
<td>BGN</td>
<td>2.1</td>
</tr>
<tr>
<td>ICAM5</td>
<td>2.1</td>
</tr>
</tbody>
</table>

**Signal transduction** (examples: intracellular signaling cascade GO:0007242, receptor activity GO:0004872, GTPase regulator GO:0030695)

| **VPH1** | 2.1 | PFC | 1.5 | PRPF4B | 1.3 | APC | >10 |
| **SNO1** | 2.1 | GAB1 | 1.5 | CENTG1 | 1.3 | KRAS | >10 |

*EMBL-EBI*
Relations in Biology for complex querying

- OBO Relation Ontology contains small set of commonly used relations in biology  [http://obofoundry.org/ro/](http://obofoundry.org/ro/)
- is-a
- part-of
- proper-part-of (as part-of except subject and object are distinct)
- derives-from
- located-in
- contained-in
- adjacent-to
- preceded-by
Relations in Biology

• When we build these models we can begin to tie together data and ask new questions
• Which diseases affect the thoracic cavity?
• Which organs are contained therein?
• Which types of cells might be affected?
Where to stop?

- **Know your use cases**

Images: Wellcome Trust & NIH
Summary

• Wide range of use cases in life sciences
• Lot of ontologies needed to meet them
• Some ontologies are seen as community standards – a mark of maturity
• We don’t have everything (and may never) and some that we do have are not up to task
• But we do have enough to do some useful things with
• Pick your poison