Introduction to BioMedical Ontologies #3: Anatomy of an Ontology Annotation, part 2

Script by Jennifer R. Smith

This video is the third in a series of tutorials about biomedical ontologies. In the first video, we talked in general about what an ontology is. In the second we looked at where ontology annotations come from and how you can differentiate between annotations made on the basis of experimental evidence and those based strictly on computational evidence.

As in our previous videos, we will primarily use the Gene Ontology or GO as our example because of its widespread use.

Obviously, the key part of an ontology annotation is the term itself. In a GO annotation, the term is what tells you what a gene or gene product is doing, where it resides in the cell or what process it's involved in. Each ontology term can also be represented by a unique, ontology-specific identifier such as a GO number for the gene ontology or a PW number for RGD's pathway ontology, and you will occasionally see an ontology ID instead of, or in addition to, the ontology term.

In some circumstances it is necessary to modify or limit the meaning of an ontology term to express what the research shows. The gene ontology uses term "qualifiers" to do that. For instance, data from one reference shows that the cell adhesion molecule Cadm2, also known as Necl-3, is localized to axons in the cerebellar white matter whereas another reference specifically demonstrates that it is not present in the axons of neurons of the dorsal root ganglion. To express this, at the Rat Genome Database, Cadm2 has been annotated to both "axon" and "NOT axon".

For more information on the use of the "NOT" qualifier as well as information on the other two GO qualifiers, contributes_to and colocalizes_with, see the "annotation conventions" page on the GO Consortium website (http://www.geneontology.org/GO.annotation.conventions.shtml)

If you've spent any time looking at gene ontology annotations at database websites like RGD and Entrez Gene, you've probably noticed a column labeled "Evidence" filled with an alphabet-soup-like list of letters. These are "Evidence Codes". An evidence code is a two to four letter abbreviation, or a phrase, which is assigned to an ontology annotation on the basis of the type of evidence that supports the association between the term and the gene, protein or other data object. In part one of Anatomy of an Ontology Annotation, we mentioned that annotations could be divided into two general types--those based on experimentation and those which are solely computational--and we showed that the way to tell which of these categories an annotation falls into is via the "evidence code".

The only evidence code that we talked about in detail in Part 1 was "IEA" or "inferred from electronic annotation". This code tells you that the annotation is based strictly on computation and hasn't been reviewed or verified by anyone. Because GO has a total of 17 different evidence codes, we won't try to cover all of them in detail here. Instead, we'll only discuss a few of the more commonly used ones to give you a feel for what an evidence code represents. Information on where to find out more about the GO evidence codes will be presented at the end of this segment as well as at the end of the video.

The experimental evidence codes are designed to give you an idea of what kinds of experiments were used to demonstrate the relevant function, process or subcellular localization. For instance, if I were to use a fluorescently labeled antibody to show the subcellular localization of a protein, any annotations made regarding that localization would use the evidence code IDA or Inferred from Direct Assay. IDA can also be used for experiments such as expressing a cloned DNA for an enzyme and directly assaying the function of the resulting protein.

On the other hand, suppose a researcher were to interfere with the normal function of a gene by reducing its expression or by inhibiting the function of its protein product. In that case, an annotation for the function, process or cellular component of that gene would be said to be "Inferred from a Mutant Phenotype", and the evidence code used would be IMP. In this case, "mutant" is used to designate a general "non-normal" phenotype rather than necessarily referring to an actual nucleotide mutation.

When a gene or gene product has a GO term for a binding activity associated with it, often the evidence code that's used is IPI or Inferred from Physical Interaction. This means that an assay which specifically demonstrates the interaction, such as a classical binding assay, copurification or a co-immunoprecipitation was used. In such cases, the annotation often includes an identifier for the "binding partner", such as a sequence ID for a gene product, or an ID from a database like PubChem or ChEBI (the Chemical Entities of Biological Interest database) for molecules like ATP and metal ions. These IDs are often displayed as part of the evidence code. Alternatively, they can be located in a separate column or field sometimes referred to as the "WITH" field meaning that the binding activity is "inferred from its demonstrated physical interaction WITH this binding partner".

The use of evidence codes like IDA, IMP and IPI means that experiments were done using that gene or gene product from that species. But suppose I wanted to make an annotation on a rat gene based on experiments done with the same gene from mouse, or on experiments done using a similar, but not the same, rat gene. In those cases, the code used would be one of four “inferred from similarity” evidence codes. Which of the four is used depends on what kind of similarity the two genes or gene products display. For example, if the two genes are known orthologs of each other as is the case with many rat, mouse and human genes, the evidence code ISO or "Inferred from Sequence Orthology" is used. In these cases the "WITH" field contains the ID of the gene or gene product on which the experiments were originally done and to which the annotated gene has similarity.

Hopefully, this gives you an idea of what lies behind the GO evidence codes. For more information about when and how each code is used, check out the Gene Ontology Consortium's evidence code documentation at geneontology.org. Also, since evidence codes comprise their own controlled vocabulary, you can explore the Evidence Code Ontology using the National Center for Biomedical Ontology's BioPortal tool at bioportal.bioontology.org.

Now that we've seen how annotations for the Gene Ontology are organized, what about other ontologies? Annotations for other ontologies can be structured in much the same way, or they can be quite different. For instance, the Rat Genome Database uses ontologies to express information about disease and phenotype associations. The structure of these annotations similar to that of GO annotations, but they use some alternative evidence codes.

REMOVED FROM FINAL VIDEO:
[One of these is IAGP, or Inferred from Association between (of) Genotype and Phenotype. This is used when a genetic alteration such as a single nucleotide polymorphism, a mutation or a copy number variant is found to be linked to a particular phenotype. Annotations include a "Notes" field that can contain specific information about the genetic alteration when that data is available. In this example (Lepr, RGD:3001, DO annotation for Obesity, ref RGD:729297), a missense mutation was found in the leptin receptor gene of the Zucker fatty rat resulting in the obese phenotype and by extension, the disease Obesity.]

So far we have talked about cases in which the annotations include more components than just the term. There are cases, however, where ontology terms are used primarily as tags to organize data and facilitate searching. For instance, the Stanford Tissue Microarray Database uses the NCI Thesaurus to tag images of tissue arrays. To find relevant image data, users can either browse a diagram of the ontology tree to find terms linked to the samples they are interested in, or they can search by one or more ontology terms. In this case, I browsed down the tree to find the term Pancreatitis and located 8 images tagged with that term. Because the term "pancreatitis" is the medical diagnosis associated with the tissue, no added data such as an evidence code is needed.

So here's a recap of what we've talked about:

- Although the ontology term itself is the most important part of an ontology annotation, sometimes there are other informative components that are presented.
- When a Qualifier is present, it limits or modifies the meaning of the ontology term.
- The Evidence Code gives you an indication of what kind of experiment was done to support the assignment of an ontology term to a data object such as a gene, and whether the experiments were done with material from the species being annotated or another species.
- In some situations additional information such as data about a binding partner or the gene or gene product on which the original experiments were done, is needed. In these cases, this data is either shown as part of the evidence code or is included in a separate "WITH" field.
- In addition to formal annotations such as those used for the Gene Ontology, ontology terms can be used as tags to organize data and facilitate searching.

For more information:

Rat Genome Database (RGD):
http://rgd.mcw.edu

The Gene Ontology Consortium (GOC):
http://www.geneontology.org/

GO Annotation Conventions:
http://www.geneontology.org/GO.annotation.conventions.shtml

GO Evidence Codes:
http://www.geneontology.org/GO.evidence.shtml

The National Center for Biomedical Ontology (NCBO)
http://www.bioontology.org/

NCBO's BioPortal:
http://www.bioontology.org/BioPortal

NCBO's Evidence Code Ontology Browser:
http://bioportal.bioontology.org/virtual/1012

NCBI's Entrez Gene:
http://www.ncbi.nlm.nih.gov/sites/entrez?db=gene

Mouse Genome Informatics (MGI):
http://www.informatics.jax.org/

Stanford Tissue Microarray Database (TMAD)
http://tma.stanford.edu/cgi-bin/home.pl