# Introduction to BioMedical Ontologies #2: Anatomy of an Ontology Annotation, part 1

Script by Jennifer R. Smith

This video is the second in a series of tutorials about biomedical ontologies.

In the first video in the series, we talked about what ontologies are. For instance, we discussed the fact that an ontology is a tool that can be used to standardize and define the terms we use to express information. That way. everyone who talks about that information is speaking the same language.

So now that you know a little about ontologies in general, let's talk how the use of ontologies can help you as a researcher.

Suppose you need information about a particular gene of interest such as what it does, if it's part of a pathway or associated with a disease or phenotype. Or suppose that you've done some microarray experiments. Now you have a list of genes and need to know if they share any attributes.

How do you find that out? You could go to the literature—search PubMed or Google Scholar. But for some genes you could easily come up with hundreds of papers that you would need to sift through to try to find the data you need. And what about genes where no one has done any experimentation—is there any way to find out what those genes might do?

Ontology annotations provide a summary of what is known about a gene or gene product. For example, examining the Gene Ontology annotations attached to a gene will give you information about what its function is, what biological processes it's involved in and where it resides in the cell. In addition, the Rat Genome Database provides annotations that tell you what diseases and phenotypes a gene is associated with and what pathways it participates in.

There are several ways by which this information is attached to gene records. The first is through the process of "manual curation". Scientific curators read papers, extract the applicable data and attach that data to gene or protein records through the use of ontology annotations. Let's use the gene "regulator of G-protein signaling 9" or Rgs9 as an example. The literature for that gene includes this paper: Brain-specific RGS9-2 is localized to the nucleus via its unique proline-rich domain. Looking at the results section, you'll find a figure showing that the protein can be localized throughout the cell or just in the nucleus. At RGD, a curator could write a note in the gene record saying that the protein could be found throughout the cell or just in the nucleus. But then a person who used the search term "cytoplasm" would not find that gene because the note didn't contain that exact word. If, however, the curator used the ontology terms "cytoplasm" and "nucleus" to describe the location of the protein, Rgs9 would come up in such a search. As you can see here, that is exactly what was done. This is the list of gene ontology terms which RGD has associated with the rat gene Rgs9, and you can see that it includes both terms.

Looking more closely at these annotations, you can see that there are some letters under the heading "Evidence", and we'll talk about those in a minute. The other two items in these annotations are a "Source"—RGD—meaning that the annotations were made by curators at the Rat Genome Database, and a number under the heading "Reference". If you click on that number, you will see that the link takes you to the abstract for the paper that we started this example with. Every annotation that is manually curated from a published paper will provide a link to that paper. That way, if you want to look at the data in more detail, you have easy access to the original article.

This is an example of two ontology annotations that were made based on experimental evidence. We know that because we started with the paper, but how can you tell which annotations from a long list are based on experimentation without checking every single reference? The answer lies in the abbreviations, or "codes", listed under "Evidence". We'll talk about evidence codes in more detail in the next video, but for now, we'll split the codes into two major groups—experimental and computational. Taking rat genes as an example, the group based on experimental evidence would include annotations made to rat genes based on experiments done using the rat gene or protein, as well as experiments done on the corresponding gene in mouse or human. In the latter case, it is inferred that the rat gene does the same thing based on its similarity to the gene in the other species.

But then, what about genes that no one has studied? Is there any way to get information about those genes? For many genes that don't have experimental data, predictions of function, subcellular localization and so forth can be made based on computer algorithms that look at general similarities between genes or gene products, such as the presence of conserved domains in the protein sequence. This RGD gene (LOC502274) is one such example. *[PLEASE NOTE: The example given here is different than the one shown in the video due to the fact that since the video was produced the ortholog assignment for the video example gene—previously LOC303823, now Map3k13—was changed and new "ISO" annotations were added. That gene no longer has any IEA annotations. A good example of the often changing nature of this type of data without any experimental confirmation.]* Although no research has been done on this gene or on the protein derived from it, it has GO annotations. In these cases, the evidence code is "IEA". This stands for "inferred from electronic annotation". That means that it is solely based on computational rather than experimental evidence and that the evidence has not been reviewed or verified by anyone. In this case, the annotation also tells you that, for instance, the assignment of the term "ATP binding" is inferred from the fact that the protein contains the core protein kinase (catalytic) domain. Because other proteins that have this domain have been shown to bind ATP, we infer that this one might also bind ATP.

Because IEA annotations are calculated by computer, rather than being based on wet-lab experimentation, they should be used as pointers to possible functions rather than absolute proof of those functions.

To summarize, here's what we've talked about:

- As you know if you've spent any time doing so, finding out information about a gene can be very time consuming if you have to do all the research yourself.
- Fortunately, curation teams at databases such as the Rat Genome Database, Mouse Genome Informatics and the UniProt Knowledgebase are already looking at papers and extracting useful facts from them in the form of ontology annotations.
- Ontology annotations succinctly capture information about a gene's function, subcellular localization and so forth.
- Ontology annotations fall into two general categories—those based on experimental evidence and those based strictly on computational evidence.
- And finally, you can use the evidence code of an annotation to distinguish which of these two categories the annotation falls into.

Actually, there is more information in the evidence code than just whether the evidence is experimental or computational. If you are interested in learning more about the gene ontology evidence codes, including what they mean and when they are used, part 2 of "Anatomy of an Ontology Annotation" will delve into more detail on that subject. In addition, that video will talk about what the other two components of ontology annotations, the qualifier and the "WITH field", are used for as well as what important information they contain.

**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/

The UniProt Knowledgebase (UniProtKB):
http://www.uniprot.org/uniprot/

About mapping InterPro domains to Gene Ontology terms:
http://www.ebi.ac.uk/GOA/InterPro2GO.html