GPAD/GPI: Next generation file format for GO annotations

Rama Balakrishnan[1], Chris Mungall[2], Heiko Dietze[2], Jane Lomax[3], Tony Sawford[3], The GOC Project

[1]Department of Genetics, School of Medicine, Stanford University, Stanford, CA 94305, USA, [2]Lawrence Berkeley National Laboratory, 1 Cyclotron Road, Berkeley, CA 94720, USA,
[3]EMBL-European Bioinformatics Institute, Hinxton, Cambridge, United Kingdom.

The Gene Ontology Consortium (GOC) is a community-based bioinformatics project that classifies gene product function through the use of structured controlled vocabularies. A fundamental application of the Gene Ontology (GO) is in the creation of gene product annotations, evidence-based associations between GO definitions and experimental or sequence-based results. Traditionally, GO annotations are recorded and supplied in a standard tab-delimited file format called the Gene Associations File (GAF, http://www.geneontology.org/GO.format.annotation.shtml).

Over the years, both curators and the community have found the need to capture/express more details for an annotation and, with increases in the amount of data, the number of annotations is having an impact in the size of the GAF file.This new file format system is called Gene Product Association Data (GPAD)/Gene Product Information (GPI) format (http://wiki.geneontology.org/index.php/Final_GPAD_and_GPI_file_format).

This new system is designed to normalize data by separating the gene product/gene information from the annotation data. Data related to gene products--symbol, name, synonyms, taxon--can be submitted, updated and maintained separately in the GPI file, while the annotation details such as GO IDs, evidence codes, references, annotation extension are stored in the GPAD file. This allows users to supply annotations for unmapped loci, supports the use of the granular Evidence Code Ontology (ECO) terms instead of the three letter GO evidence codes, relationship between the gene product and the GO term, date when all the annotations for a gene was completed or reviewed, to name a few advantages. Annotating groups can opt to supply annotations in the traditional 17 column GAF file or in the GPAD file format. The GOC provides scripts to convert between the two file formats in addition to a validator to validate the files.

The Gene Ontology Consortium is funded by NIH NHGRI HG 002273 to PIs Judith A Blake, J. Michael Cherry, Suzanna Lewis, Paul W Sternberg and Paul Thomas.

The Gene Ontology (GO) Project aims to:

1. Maintain and develop controlled vocabulary of terms (GO Terms) for describing gene products
2. Annotate data (association of GO Terms to gene products)
3. Assimilate and disseminate annotation data
4. Provide tools for easy access of all the data

Contact GOC:
http://www.geneontology.org/GO.contacts.shtml
go-helpdesk@geneontology.org

Disadvantages of the Gene Association File (GAF) format

1. Large denormalized file
2. Combined representation of gene product data and annotations leads to repetition
3. No system to represent gene product metadata for unannotated genes
4. Requirement to maintain backwards compatibility makes it harder to introduce enhancements

Advantages of the Gene Product Association Data (GPAD) and Gene Product Information (GPI) file formats

1. Smaller normalized files
2. Gene product data is separated from the annotations, reduces repetition
3. The GPI file can provide information on unannotated gene products
4. Supports use of the Evidence Code Ontology (ECO)
5. A relationship between the gene product and the GO term can be specified
6. Date curation of a gene product was completed can be captured
7. A variety of annotation properties can be recorded, such as curator name' or annotation identifier
8. Flexible for both GO and non-GO curation, such as phenotype annotation

Example Rows from the GPAD with new features highlighted (not all columns are shown)

DB	DB Object ID	Qualifier	GOID	Reference	Evidence	Date Assigned	Assigned by	Annotation Extension	Annotation properties
UniProtKB	P00546	enables	GO:0004674 (protein serine/ threonine kinase activity)	PMID:24319056	ECO:0000314	20131220	SGD	has_direct_input (SGD:S000000520)	go_evidence=IDA| id=2113694141| curator_name= Joe Smith
IntAct	EBI-8874189	part_of	GO:1902515 (thioredoxin-disulphide reductase complex)	PMID:15144954	ECO:0000314	2014-01-01	IntAct		go_evidence = IDA | id=123456789 | curator_name = Joe Smith
IntAct	EBI-8874189	involved_in	GO:0006739 (NADP metabolic process)	PMID:15144954	ECO:0000314	2014-01-01	IntAct		go_evidence = IDA | id=123456789 | curator_name = Joe Smith

Example Rows from the GPI file with new feature(s) highlighted (not all columns are shown)

DB	DB Object ID	DB_Object_Symbol	DB_Object_Type	DB_Object_Taxon	DB xfref	Gene_Product_Properties
UniProtKB	P00546	CDC28	protein	559292	SGD:S000000364	go_annotation_complete= 20080421
IntAct	EBI-8874189	TrxB complex	protein complex	83333		go_annotation_complete= 2014-01-02

Note: GAF files will continue to be supported, but we are encouraging annotating groups to move to the new file formats. A conversion script between GAF to GPAD/GPI and back is available.