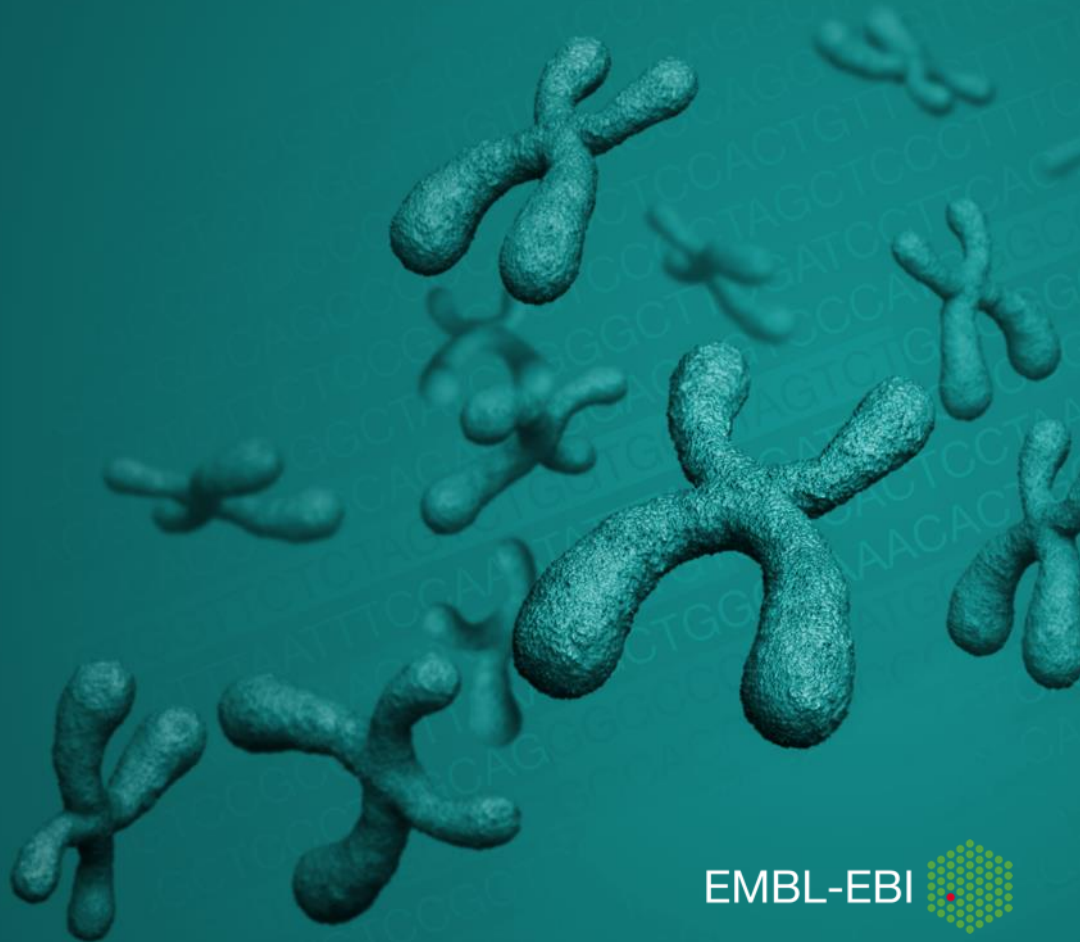# Gene Ontology Annotation Extensions

Increasing the expressivity of GO

Prudence Mutowo Ph.D.

EMBL-EBI

The Gene Ontology scope

- Molecular Function, e.g. *kinase activity, opsin binding*

- Biological Process, e.g. *cell cycle, limb development*

- Cellular Component, e.g. *nucleus, SMAD protein complex*

GO annotation practice

- Each GO annotation can pair only a single gene product with a single GO term

- Limits the complexity of information that can be captured

- Annotations cannot be linked

Cannot capture:

- Targets of a process/function,

    e.g. hcn1 is phosphorylated by the serine/threonine kinase cdc2

- Localization dependency,

    e.g. protein 'x' acts as a protein kinase in the nucleus but not the cytoplasm

- Enzyme substrates or products,

    e.g. epsin-1 ubiquitination

- Processes involved in other processes,

    e.g. localization of protein to nucleus in response to oxidative stress

GO annotation practice

- GO terms can be pre-composed for some of these, e.g.

"protein localization to nucleus in response to oxidative stress"

But, we shouldn't attempt to make a term for *everything*

GO terms will always be a *subset* of total set of possible descriptions

Annotation Extensions

New!

Allows curators to combine GO terms with:

- other GO terms

- terms from external ontologies

- database identifiers representing genes or their products

That is, post-composition at the time of annotation

EMBL-EBI

Semantics

Each annotation extension consists of a relationship-identifier pair, e.g.

occurs_in(CL:0000066)

Relationship   **Identifier**

Each relationship has usage restrictions

- The primary GO terms (Col.5 of GAF) that can be used with the relationship

- The identifiers that may be used in the extension

Note: the extension provides the relationship of the GO term to other entities
NOT of the gene product to other entities

Semantics - example

e.g. *occurs_in*

- can only be used when annotating with GO:molecular function or GO:biological process

- is restricted to use with identifiers from cell type ontology (CL), anatomical ontology (UBERON), plant anatomical ontology (PO), C. elegans cell or anatomy ontology (WBbt) or GO:cellular component

AMACR   GO:0008111   IMP   PMID:10655068   occurs_in(CL:0000057)

where GO:0008111 is alpha-methylacyl-CoA racemase activity (molecular function)

and CL:0000057 is fibroblast (cell type ontology)

EMBL-EBI

Semantics

- Over 40 **relationships** have been defined so far – more can be added
- Legal **identifiers** currently include:
  - ChEBI
  - Cell Type (CL)
  - Tissue Type (Uberon)
  - Sequence Ontology (SO)
  - Protein Ontology (PR)
  - Protein Complexes (MI)
  - Plant Ontology (PO)
  - Zebrafish development stages (ZFS)
  - *C. elegans* cell and anatomy ontology (WBbt)
  - Protein modifications (MOD)
  - Gene Ontology (GO)
  - Gene/protein identifiers (Ensembl, UniProt, PomBase etc.)

  – more can be added
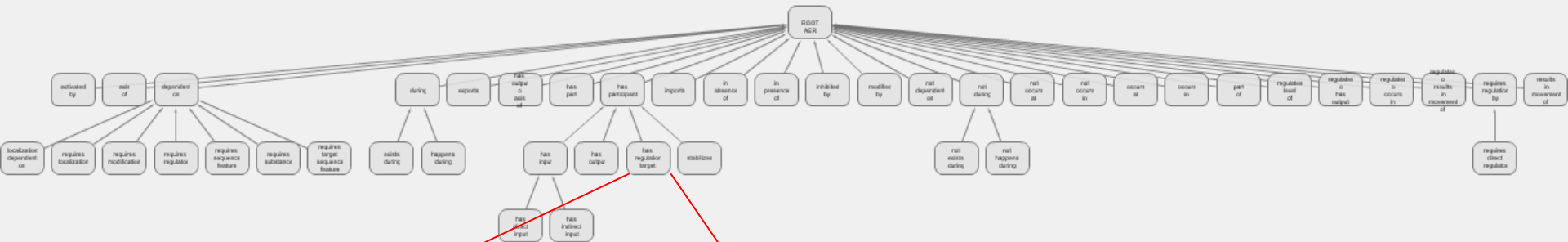
EMBL-EBI

Graphical visualisation of relations

EMBL-EBI

Independent and conjunctive extensions

Independent extensions (OR operator) are used to make separate statements about the same GO term, e.g. where a gene product can perform its activity in multiple places or under various conditions

"Pap1 is a transcription factor that regulates transcription of several oxidative stress response genes, including *trr1*, *pmd1*, *obr1*, *bfr1* and *ctt1*"

Pap1   GO:0045893 (positive regulation of transcription) IDA   PMID:9585505

Annotation extensions are pipe separated:

has_regulation_target(PomBase:SPBC3F6.03)|has_regulation_target(PomBase:SPCC663.03)|
has_regulation_target(PomBase:SPAC3C7.14c)|has_regulation_target(PomBase:SPCC18B5.01c)|
has_regulation_target(PomBase:SPCC757.07c)

This is equivalent to making five separate annotations

EMBL-EBI

Independent and conjunctive extensions

Conjunctive extensions (AND operator) are used to make linked statements about the same GO term where two or more extensions apply simultaneously, e.g. when a gene product acts upon a second gene product during the course of another process

Example 1:

"Upon oxidative stress, Sty1 regulates the nuclear localization of Pap1"

Sty1   GO:0034504 (protein localization to nucleus)   IMP   PMID:9585505

Annotation extensions are comma separated:

happens_during(GO:0034599),has_input(UniProtKB:Q01663)

Where GO:0034599 is 'cellular response to oxidative stress' and Q01663 is Pap1

Independent and conjunctive extensions

Conjunctive extensions

Example 2:

"TMEM115 is located in the nucleus that is part of an epithelial cell that is part of the cervix epithelium"

TMEM115   GO:0005634 (nucleus)   IDA   PMID:17973242

Annotation extensions are comma separated:

part_of(CL:0000066),part_of(UBERON:0004801)

Where CL:0000066 is 'epithelial cell' and UBERON:0004801 is 'cervix epithelium'

Other examples

1. When a protein both positively and negatively regulates a process/activity under certain conditions, e.g. when bound to something, when located in a particular cell or component, etc.

2. When a protein is located in secretions, such as tears or saliva; currently have no GO terms for these, only 'extracellular space'.
Could add in an extension using Uberon terms, e.g. 'lacrimal gland secretion' or 'saliva'

Where can I get them?

- Download
  - http://geneontology.org/GO.downloads.annotations.shtml
    - Groups providing: UniProt, IntAct, MGI, PomBase, DictyBase, AgBase, WormBase
- Search and Browsing
  - Cross-species
    - AmiGO 2 – http://amigo2.berkeleybop.org
    - QuickGO (next year) - http://www.ebi.ac.uk/QuickGO/
  - MOD interfaces
    - PomBase – http://pombase.org

EMBL-EBI

Display of annotation extensions in PomBase

pap1 (SPAC1783.07c)

GO:0036091	positive regulation of transcription from RNA polymerase II promoter in response to oxidative stress		IMP	Castillo EA et al. (2002)	2
	regulates pmd1		IMP	Toone WM et al. (1998)	
	regulates bfr1		IMP	Toone WM et al. (1998)	
	regulates ctt1		IMP	Toone WM et al. (1998)	
	regulates pgr1			97)	
	regulates trr1	Independent extensions		l. (1998)	
	regulates obr1		IMP	Toone WM et al. (1998)	

GO:0001077	RNA polymerase II core promoter proximal region sequence-specific DNA binding transcription factor activity involved in positive regulation of transcription				26
	regulates sod1		IMP	Mutoh N et al. (2002)	
	regulates gst1	Conjunctive extension	IMP	Kim HG et al. (2004)	
	regulates obr1		IMP	Toda T et al. (1992)	
	during regulation of transcription from RNA polymerase II promoter in response to oxidative stress, regulates hsr1		IMP	Chen D et al. (2008)	
	during regulation of transcription from RNA polymerase II promoter in response to oxidative stress, regulates pof14		IMP	Tafforeau L et al. (2006)	

Note, PomBase has replaced annotation extension relation names
with more human-readable versions in their web display

EMBL-EBI

Display of annotation extensions in AmiGO2

Curation tool support

- Supported in
  - Protein2GO (UniProt, DictyBase, WormBase, AgBase, SGD)
  - CANTO (PomBase)
  - MGI curation tool

Curating annotation extensions in Protein2GO

Curating annotation extensions in Protein2GO

Curating annotation extensions in Protein2GO

Annotation Extension

Statement Group

Relation subset	Relation	Database	Identifier	
sequence_or_complex	has_direct_input	UniProtKB	Q4VCS5	AMOT (Homo sapiens): Angiomotin

Statement Group

Relation subset	Relation	Database	Identifier	
sequence_or_complex	has_direct_input	UniProtKB	P12345	GOT2 (Oryctolagus cuniculus): Aspartate aminotransferase, mitochondrial

Conjunctive extension

Independent extensions

Add new statement group

Copy to scratchpad   Paste from scratchpad: (empty)

Show annotation_extension relations ontology in QuickGO

Current value:
has_direct_input(UniProtKB:Q4VCS5)|has_direct_input(UniProtKB:P12345)

OK   Cancel

transcription

EMBL-EBI

Curating annotation extensions in Protein2GO

Are annotation extensions useful?

Need to have:

- a larger sample set

- more curators/groups contributing annotation extensions

- GO analysis tool providers to include annotation extensions

GO annotations with extensions can give an improved level of detail and interconnection, supporting more sophisticated querying and analysis

e.g. give me all the human proteins that are involved in blood vessel remodeling during retina vasculature development in the camera-type eye.

e.g. give me all the mouse proteins that have protein kinase activity in kidney cells

A method for increasing expressivity of Gene Ontology annotations using a compositional approach

EMBL-EBI

Resources

- Introduction to annotation extensions:
  http://www.geneontology.org/GO.annotation.extension.shtml

- Usage examples for each relationship (not complete):
  http://wiki.geneontology.org/index.php/Annotation_usage_examples_for_each_annotation_extension_relation

- Interactive graphical visualisation of relationships:
  http://www.ebi.ac.uk/QuickGO/AnnotationExtensionRelations.html

- OBO Foundry – for finding other ontology IDs, e.g. CL or UBERON:
  http://www.obofoundry.org/

- Ontology Lookup Service:
  http://www.ebi.ac.uk/ontology-lookup/

- Annotation Extension Relations OBO file:

http://viewvc.geneontology.org/viewvc/GO-SVN/trunk/ontology/extensions/go_annotation_extension_relations.obo?view=log

Acknowledgments

- Chris Mungall, Tony Sawford

- GO Consortium, model organism and UniProt curators

- PomBase developers:

  - Mark McDowell, Kim Rutherford

Reference: *Huntley,R., Harris,M., Alam-Faruque,Y., Blake,J., Carbon,S., Dietze,H., Dimmer,E., Foulger,R., Hill,D., Khodiyar,V., et al. (2014) A method for increasing expressivity of Gene Ontology annotations using a compositional approach. BMC Bioinformatics, **15**, 155.*

EMBL-EBI