

GOAT: The Gene Ontology Annotation Tool

Dr. Mike Bada

Department of Computer Science

University of Manchester

mbada@cs.man.ac.uk



Outline

- Semantic annotation and ontologies
- The Gene Ontology
- Annotation with terms from the Gene Ontology
- GOAT
- Conclusions



The Need for Semantic Annotation

- Bioinformatics has a large number of distributed, autonomous, heterogeneous resources
- The experimenter and computer need access to the knowledge within these resources
- Data need to be in a common, computationally amenable form



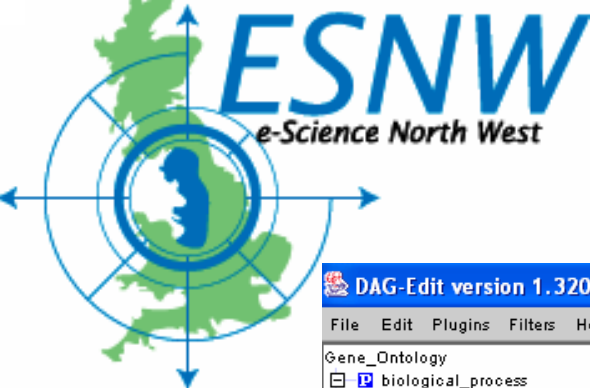
Ontologies

- An ontology is a set of terms, relationships, and definitions that capture a community's understanding of a domain
- Terms represent the concepts of a domain, which are linked by relationships; these constitute a controlled vocabulary
- These terms augment natural-language annotation and can be more easily processed computationally



The Gene Ontology (GO)

- Is currently comprised of over 15,000 terms representing molecular functions, biological processes, and cellular components
- Is arranged as three orthogonal, structurally unlinked subontologies
- Enables a common understanding between databases and between model organisms
- Has been used extensively in a number of prominent biological databases



A Fragment of GO

DAG-Edit version 1.320

File Edit Plugins Filters Help

Gene_Ontology

- [-] biological_process
 - [+] behavior
 - [-] biological_process unknown
 - [+] cellular process
 - [+] development
 - [+] obsolete
 - [+] physiological processes
 - [-] viral life cycle
 - [+] latent virus infection
 - [-] lysogeny
 - [+] provirus integration
 - [+] regulation of viral protein levels
 - [+] viral infectious cycle
 - [+] viral spread within host
 - [+] viral transformation
 - [+] viral transmission
 - [+] virus-host interaction
- [-] cellular_component
 - [+] cell
 - [-] cellular_component unknown
 - [+] extracellular
 - [-] immunoglobulin complex
 - [+] immunoglobulin complex, circulating
 - [+] immunoglobulin complex, membrane bound
 - [+] obsolete
 - [+] unlocalized
 - [+] virion
- [-] molecular_function
 - [+] anticoagulant activity
 - [+] antifreeze activity
 - [+] antioxidant activity
 - [+] apoptosis regulator activity
 - [-] binding activity
 - [+] acyl binding activity
 - [+] amino acid binding activity
 - [-] antigen binding activity
 - [+] peptide antigen binding
 - [+] endogenous peptide antigen binding
 - [+] exogenous peptide antigen binding
 - [+] toxin binding activity
 - [+] boron binding activity
 - [+] calcium oxalate binding activity
 - [+] carbohydrate binding activity
 - [+] cofactor binding activity

Find terms

ID contains Find

Search all terms
 Search children of selection
 Search selection
 Case sensitive search

No search performed

ID: **GO:0042605**

Term name: peptide antigen binding

Definition:

Text: Dbxrefs:

Synonyms:

Select a synonym from the list to edit it, or press add to create a new synonym

General DbXrefs:

Select a dbxref from the list to edit it, or press add to create a new dbxref

Comment:

DAG Viewer

Gene_Ontology

- [-] molecular_function
 - [+] binding activity
 - [+] antigen binding activity
 - [+] peptide antigen binding
 - [+] peptide binding activity
 - [+] peptide antigen binding
 - [+] defense/immunity protein activity
 - [+] antigen binding activity
 - [+] peptide antigen binding



Annotation with GO Terms

- Annotator must rely upon his/her domain expertise and the usability of the annotation tool
- S/he must find terms among the 15,000+ terms
- Unconstrained entry of terms may lead to inconsistent or nonsensical descriptions of gene products
- GOAT dynamically determines which terms are most relevant and offers these subsets as options



GONG and GOAT

- GOAT relies upon GONG
- Goal of GONG is to convert GO into a more formal, Description Logic representation and enrich its content
- Description Logics provide unambiguous semantics of a model and come with reasoning support
- Specifically, DAML+OIL is being used



Adding Associations to GO

- All GO annotations have been gathered into one database named GOA
- GOA was mined for associations between GO terms
- GO-associated databases were also manually examined for associations between GO terms and gene-product types
- These GO-term-to-GO-term and GO-term-to-gene-product-type associations were added to DAML+OIL GO



Guiding the Annotator

- GOAT seeks to reduce the potential for annotation errors and reduce the overhead of GO's size
- Our approach is to use a formal DAML+OIL GO, augmented with these associations, and a reasoner to guide the user in the annotation process
- GOAT presents subsets of the appropriate subontologies that are most likely relevant in that they have been associated with information already entered by the user

GOAT demo [-] [] [X]

name	<input type="text"/>
gene-product type	<ul style="list-style-type: none">messenger RNA (mRNA)proteinsmall nuclear RNA (snRNA)small nucleolar RNA (snoRNA)transfer RNA (tRNA)
molecular function	<input type="text"/> Add term
biological process	<input type="text"/> Add term
cellular component	<input type="text"/> Add term



An Annotation Scenario

Say the user has a specific snRNA to annota

GOAT demo

name

gene-product type

molecular function

biological process

cellular component



An Annotation Scenario

GOAT demo

name

gene-product type

- messenger RNA (mRNA)
- protein
- small nuclear RNA (snRNA)**
- small nucleolar RNA (snoRNA)
- transfer RNA (tRNA)

molecular function

biological process

cellular component

molecular function terms

RNA binding activity

- ribonuclease activity
 - endoribonuclease activity
 - endonuclease G activity
 - endoribonuclease activity, producing 5'-phosphomonoesters
 - RNA lariat debranching enzyme activity
 - ribonuclease H activity
 - ribonuclease H1 activity
 - ribonuclease III activity
 - bidentate ribonuclease III activity
 - endoribonuclease activity, producing other than 5'-phosphomonoesters
 - enterobacter ribonuclease activity
 - pancreatic ribonuclease activity
 - tRNA-intron endonuclease activity
 - exoribonuclease activity
 - exoribonuclease activity, producing 3'-phosphomonoesters
 - exoribonuclease activity, producing 5'-phosphomonoesters
 - 3'-5'-exoribonuclease activity
 - 5'-3'-exoribonuclease activity
 - exoribonuclease H activity
 - exoribonuclease II activity
 - poly(A)-specific ribonuclease activity
 - oligoribonuclease activity
 - ribonuclease E activity
 - ribonuclease G activity
 - ribonuclease MRP activity
 - ribonuclease R activity
 - ribonuclease T1 activity
 - tRNA-specific ribonuclease activity
 - ribonuclease P activity
 - tRNA-intron endonuclease activity



An Annotation Scenario

GOAT demo

name

gene-product type

- messenger RNA (mRNA)
- protein
- small nuclear RNA (snRNA)**
- small nucleolar RNA (snoRNA)
- transfer RNA (tRNA)

molecular function

biological process

cellular component



An Annotation Scenario

GOAT demo

name

gene-product type
messenger RNA (mRNA)
protein
small nuclear RNA (snRNA)
small nucleolar RNA (snoRNA)
transfer RNA (tRNA)

molecular function

biological process

cellular component

biological process terms

mRNA splicing
MAT a1 (A1) pre-mRNA splicing
mRNA cis splicing
mRNA trans splicing
splicing AT-AC intron
splicing GT-AG intron



An Annotation Scenario

GOAT demo [minimize] [maximize] [close]

name

gene-product type
messenger RNA (mRNA)
protein
small nuclear RNA (snRNA) [highlighted]
small nucleolar RNA (snoRNA)
transfer RNA (tRNA)

molecular function

biological process

cellular component



An Annotation Scenario

GOAT demo

name

gene-product type
messenger RNA (mRNA)
protein
small nuclear RNA (snRNA)
small nucleolar RNA (snoRNA)
transfer RNA (tRNA)

molecular function

biological process

cellular component

cellular component terms

spliceosome complex
major (U2-dependent) spliceosome
minor (U12-dependent) spliceosome complex



Conclusions

- *GO*-term annotation can be tedious and error-prone
- Representing *GO* as a formal DAML+OIL ontology allows for complex and efficient reasoning over the ontology
- Reasoning over *GO* augmented with mined term associations can help guide users in the annotation process by narrowing down term choices to those that are most likely relevant



Acknowledgments

University of Manchester

Robert Stevens

Chris Wroe

Kevin Garwood

Phil Lord

Daniele Turi

Carole Goble

GlaxoSmithKline

Robin McIntire



Thanks!