

Comparing knowledge resource designs to support term-level text annotation

Alicia Tribble
Jin-Dong Kim
Tomoko Ohta
Jun'ichi Tsujii



Department of Computer
Science
University of Tokyo

Plan of the Talk

- Background: Functional Requirements and KSs
- GO, Mesh and GENIA
- Problems in GENIA
- New Scheme



Plan of the Talk

- Background: Functional Requirements and KSs
- GO, Mesh and GENIA
- Problems in GENIA
- New Scheme



Knowledge resources for Biology

- What they are:
 - Structured repositories for biological knowledge
 - Dictionaries, thesauri, topic hierarchies, domain models, databases & ontologies
- What they do:
 - **Organize knowledge in the biology field** on a shared vocabulary (**GO**)
 - **Categorize text in the biology field** for document retrieval (**MeSH**)
 - **Provide text annotation for information extraction** in Biology domain (**GENIA**)



Example: using plain text to retrieve documents

The screenshot displays the PubMed search interface. At the top, the NCBI logo is on the left, and the PubMed logo with the URL www.pubmed.gov is in the center. To the right, it states "A service of the U.S. National Library of Medicine and the National Institutes of Health". Further right are links for "My NCBI", "Sign In", and "Register". Below this is a navigation bar with tabs for "All Databases", "PubMed", "Nucleotide", "Protein", "Genome", "Structure", "OMIM", "PMC", "Journals", and "Books". The search bar contains the text "Transcription" and is annotated with a red circle and an arrow pointing to the text "Search Term". Below the search bar are buttons for "Limits", "Preview/Index", "History", "Clipboard", and "Details". The "Save Search" link is to the right. Below the search bar, there are options for "Display" (set to "Summary"), "Show" (set to "20"), "Sort By", and "Send to". A red circle highlights the text "All: 323850" and "Review: 32040". Below this, the text "Items 1 - 20 of 323850" is circled in red, with an arrow pointing to the text "Matching Abstracts". The search results are listed below, with the first four items shown. Each item has a checkbox, a link to the abstract, and a "Related Articles" link. The first item is:
 1: [Muzzini DM, Plevani P, Boulton SJ, Cassata G, Marini F.](#) Related Articles
Caenorhabditis elegans POLQ-1 and HEL-308 function in two distinct DNA interstrand cross-link repair pathways. DNA Repair (Amst). 2008 May 8. [Epub ahead of print] PMID: 18472307 [PubMed - as supplied by publisher]

The second item is:
 2: [Belleau F, Nolin MA, Tourigny N, Rigault P, Morissette J.](#) Related Articles
Bio2RDF: Towards a mashup to build bioinformatics knowledge systems. J Biomed Inform. 2008 Mar 21. [Epub ahead of print] PMID: 18472304 [PubMed - as supplied by publisher]

The third item is:
 3: [Cobaleda C, Busslinger M.](#) Related Articles
Developmental plasticity of lymphocytes. Curr Opin Immunol. 2008 May 8. [Epub ahead of print] PMID: 18472258 [PubMed - as supplied by publisher]

The fourth item is:
 4: [Moses D, Drago J, Teper Y, Gantois I, Finkelstein DI, Home MK.](#) Related Articles
Fetal striatum- and ventral mesencephalon-derived expanded neurospheres rescue dopaminergic neurons in vitro and the nigro-striatal system in vivo.

Search for important concepts and exclude non-relevant strings

The screenshot shows the PubMed search results page. At the top, the NCBI logo and 'PubMed' text are visible, along with the URL 'www.pubmed.gov'. The search bar contains the query 'Transcription, Genetic'[Majr], which is circled in red. A red arrow points from this search term to a white box labeled 'MeSH Heading'. Below the search bar, there are buttons for 'Limits', 'Preview/Index', 'History', 'Clipboard', and 'Details'. The 'Display' section shows 'Summary' selected, 'Show 20' items, and 'Sort By' options. The total number of results is 'All: 38945', which is also circled in red. A red arrow points from this number to a white box labeled 'Matching Abstracts (10 x fewer)'. The results list shows four items, each with a checkbox, author names, title, journal information, and PMID. The first item is 'Johnson DJ, Johnson SA. Cell biology. RNA metabolism and oncogenesis. Science. 2008 Apr 25;320(5875):461-2. No abstract available. PMID: 18436765'. The second item is 'Riley T, Sontag E, Chen P, Levine A. Transcriptional control of human p53-regulated genes. Nat Rev Mol Cell Biol. 2008 May;9(5):402-12. Review. PMID: 18431400'. The third item is 'LeRoy G, Rickards B, Flint SJ. The double bromodomain proteins Brd2 and Brd3 couple histone acetylation to transcription. Mol Cell. 2008 Apr 11;30(1):51-60. PMID: 18406326'. The fourth item is 'Price DH. Poised polymerases: on your mark...get set...go! Mol Cell. 2008 Apr 11;30(1):7-10. Review. PMID: 18406322'. Each item has a 'Related Articles, Links' link to its right.

NCBI PubMed A service of the U.S. National Library of Medicine and the National Institutes of Health [My NCBI](#) [\[Sign In\]](#) [\[Register\]](#)

All Databases PubMed Nucleotide Protein Genome Structure OMIM BMC Journals Books

Search PubMed for "Transcription, Genetic"[Majr] **MeSH Heading**

Limits Preview/Index History Clipboard Details

Display Summary Show 20 Sort By Send to

All: 38945 Review: 3367

Items 1 - 20 of 38945 **Matching Abstracts (10 x fewer)** Page 1 of 1948 Next

1: [Johnson DJ, Johnson SA.](#) [Related Articles, Links](#)
Cell biology. RNA metabolism and oncogenesis. Science. 2008 Apr 25;320(5875):461-2. No abstract available. PMID: 18436765 [PubMed - indexed for MEDLINE]

2: [Riley T, Sontag E, Chen P, Levine A.](#) [Related Articles, Links](#)
Transcriptional control of human p53-regulated genes. Nat Rev Mol Cell Biol. 2008 May;9(5):402-12. Review. PMID: 18431400 [PubMed - indexed for MEDLINE]

3: [LeRoy G, Rickards B, Flint SJ.](#) [Related Articles, Links](#)
The double bromodomain proteins Brd2 and Brd3 couple histone acetylation to transcription. Mol Cell. 2008 Apr 11;30(1):51-60. PMID: 18406326 [PubMed - indexed for MEDLINE]

4: [Price DH.](#) [Related Articles, Links](#)
Poised polymerases: on your mark...get set...go! Mol Cell. 2008 Apr 11;30(1):7-10. Review. PMID: 18406322 [PubMed - indexed for MEDLINE]

Hierarchical relations affect document classification

NCBI PubMed A service of the U.S. National Library of Medicine and the National Institutes of Health

My NCBI [Sign In] [Register]

All Databases PubMed Nucleotide Protein Genome Structure Books

Search PubMed for **Transcription, Genetic**[Majr:NoExp]

Limits Preview/Index History Clipboard Details

Display Summary Show 20 Sort By Send to

All: **38866** Review: 3361

Items 1 - 20 of 38866

Page 1 of 1944 Next

1: [Johnson DJ, Johnson SA.](#) Related Articles, Links
Cell biology. RNA metabolism and oncogenesis. Science. 2008 Apr 25;320(5875):461-2. No abstract available. PMID: 18436765 [PubMed - indexed for MEDLINE]

2: [Riley T, Sontag E, Chen P, Levine A.](#) Related Articles, Links
Transcriptional control of human p53-regulated genes. Nat Rev Mol Cell Biol. 2008 May;9(5):402-12. Review. PMID: 18431400 [PubMed - indexed for MEDLINE]

3: [LeRoy G, Rickards B, Flint SJ.](#) Related Articles, Links
The double bromodomain proteins Brd2 and Brd3 couple histone acetylation to transcription. Mol Cell. 2008 Apr 11;30(1):51-60. PMID: 18406326 [PubMed - indexed for MEDLINE]

4: [Price DH.](#) Related Articles, Links
Poised polymerases: on your mark...get set...go! Mol Cell. 2008 Apr 11;30(1):7-10. Review.

A topic hierarchy gives more control over the results

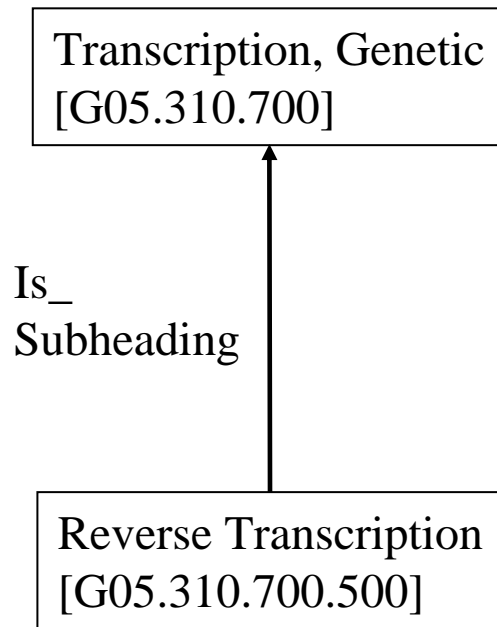
MeSH Heading	Transcription, Genetic
Tree Number	G05.310.700
Scope Note	The biosynthesis of RNA carried out on a template of DNA . The biosynthesis of DNA from an RNA template is called REVERSE TRANSCRIPTION .
Entry Term	Early Gene Transcription
Entry Term	Genetic Transcription
Entry Term	Late Gene Transcription
See Also	Gene Products, rev
See Also	Gene Products, tat
See Also	Gene Products, tax
See Also	Genes, pX
See Also	Genes, rev
See Also	Genes, tat
See Also	Trans-Activators
See Also	Transcription Factors
Allowable Qualifiers	DE ES GE IM PH RE
Entry Version	TRANSCRIPTION GENET
Previous Indexing	RNA, Messenger (1966-1972)
History Note	1973
Date of Entry	19990101
Unique ID	D014158

MeSH Tree Structures

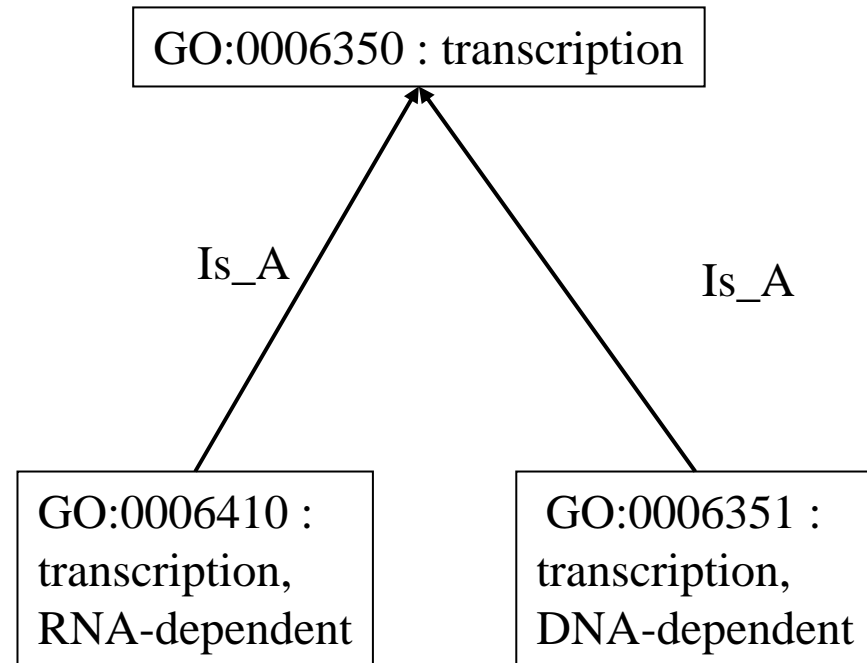
[Genetic Processes \[G05\]](#)
 [Gene Expression \[G05.310\]](#)
 [Protein Biosynthesis \[G05.310.670\]](#)
 ▶ [Transcription, Genetic \[G05.310.700\]](#)
 [Reverse Transcription \[G05.310.700.500\]](#)



Hierarchical structures differ among knowledge resources



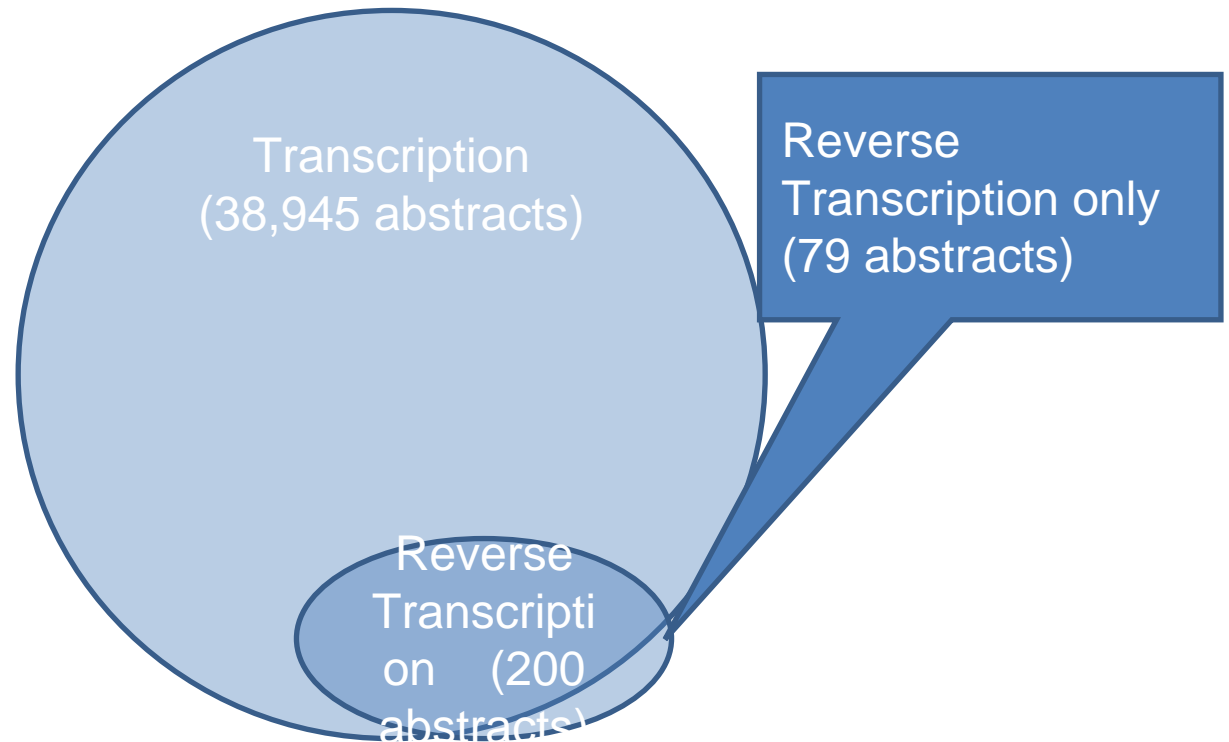
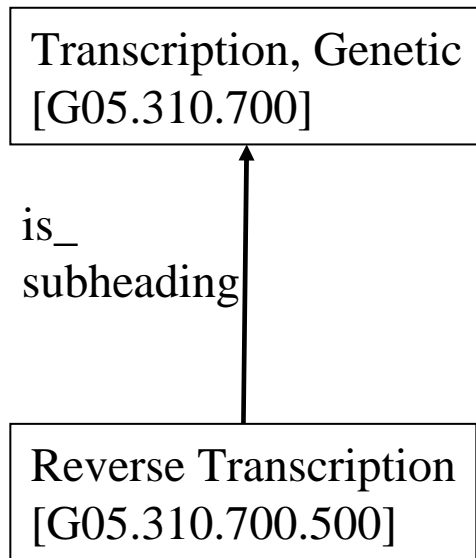
MeSH



OBO biological process

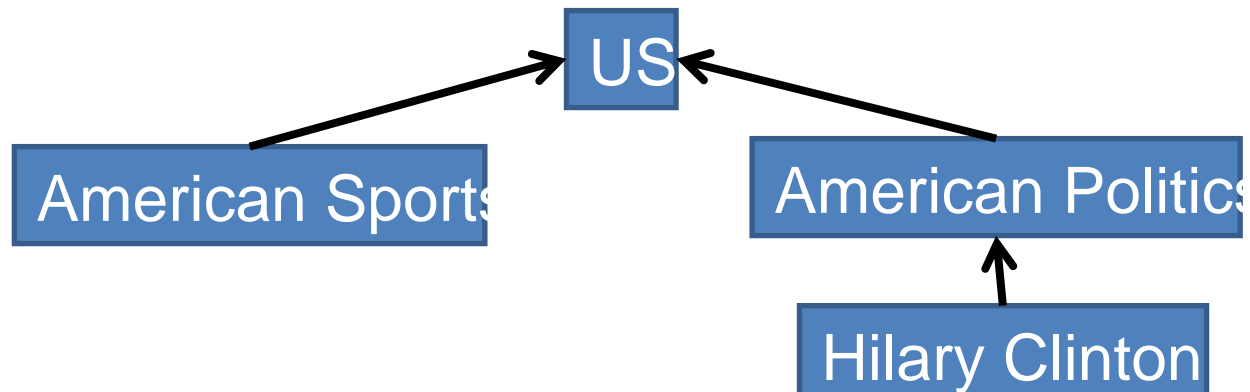


Hierarchical relations affect document classification



Hierarchical structures differ among knowledge resources

- Differences in Design:
 - “Transcription” has one child in MeSH, two in GO
- Differences in Function:
 - **MeSH “Transcription” is a topic used to sort and retrieve full documents**
 - **GO “Transcription” is a real-world event that can be associated with a set of experimental results**
 - **The hierarchies have different interpretations**



Hierarchical structures differ among knowledge resources

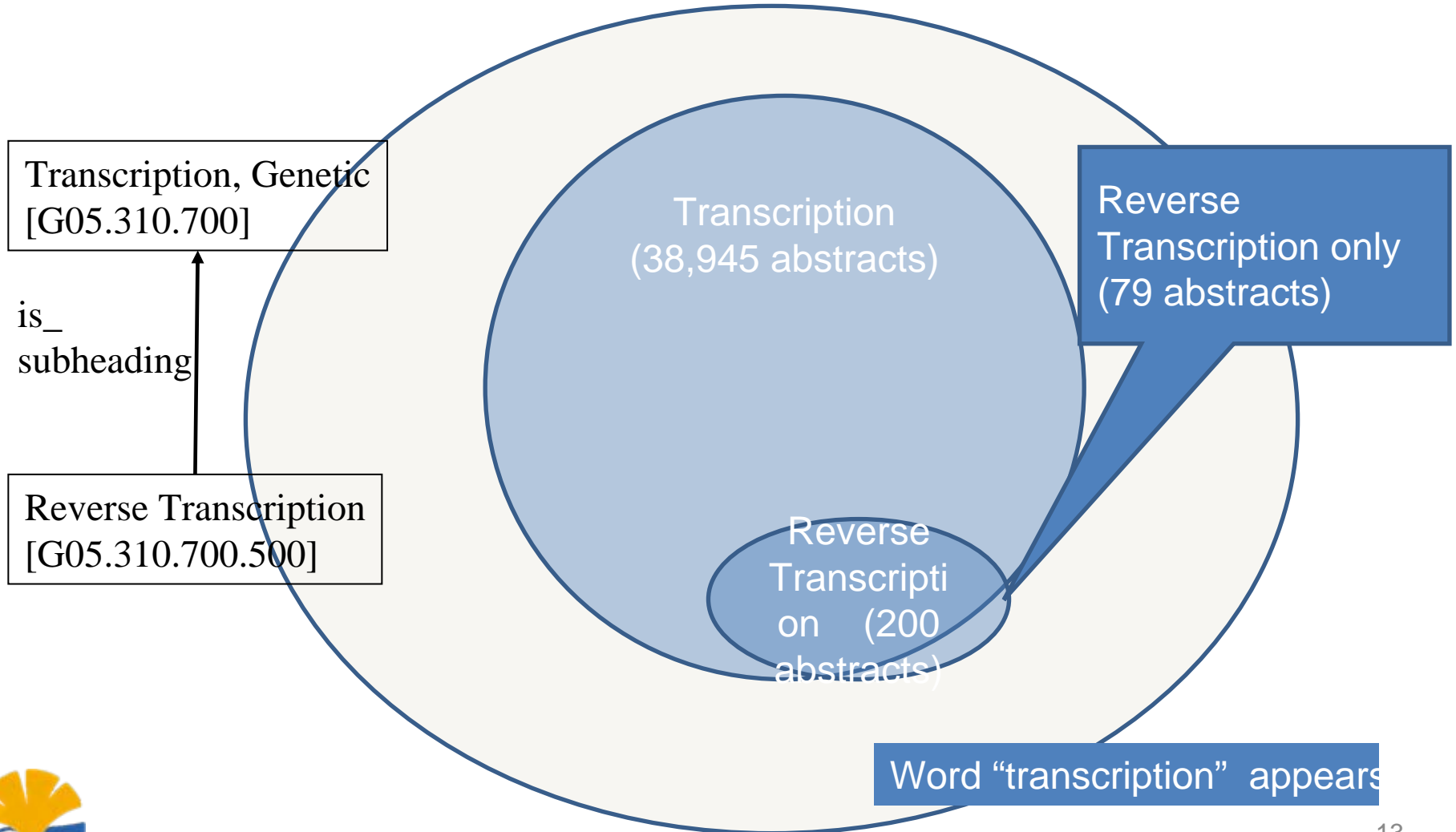
- Differences in Design:
 - “Transcription” has one child in MeSH, two in GO
- Differences in Function:
 - MeSH “Transcription” is a topic used to sort and retrieve full documents
 - GO “Transcription” is a real-world event that can be associated with a set of experimental results
 - **GENIA**

Mesh Text annotation: on text, not the biological
GO reality

Segments of text which describe biological realities



Hierarchical relations affect document classification



Units of Annotation

GO: Full papers as evidence

Mesh: Abstracts

Our focus is on vocabulary terms that can be associated with contiguous sub-sentential spans of text: *term-level annotation*.



Describing the Differences

- **Granularity:** **TEXT** ailed are the representations?

A structured knowledge resource for term-level annotation should represent terms and relations that are commonly expressed in contiguous spans of text in the target domain.

REALI

- **Interoperability:** Do the entities correspond to state-of-the-art shared knowledge of the domain?

A structured knowledge resource for term-level annotation should represent consistently-defined relations among terms that bear meaningful correspondences with other ontological knowledge resources of the target domain.



Plan of the Talk

- Background: Functional Requirements and KS
- GO, Mesh and GENIA
- Problems in GENIA
- New Scheme



Gene Ontology: Annotation

- Annotation-related uses:
 - **Controlled vocabulary used to label biology databases, so that experimental results from diverse laboratories can be automatically compared (Camon, Magrane et al. 2004).**
- Annotation example: The sentence shown is an excerpt from PubMed document PMID:12458670. GO annotation is given as a database entry that can be paraphrased as:

Drug binding, GO term 0008144, occurs between Alb and the chemical with CHEBI ID:28939, based on experimental evidence type IPI described in the paper with PubMed ID 12458670.

Covalent binding between N-acetyl-L-cysteine (NAC) and albumin was evaluated kinetically by conducting in vitro experiments.

GO ID: 0008144; Alb; CHEBI:28939; PMID:12458670; IPI



Gene Ontology: Granularity & Interoperability

- Interoperability is Strong:
 - Open development model with diverse contributors
 - Adheres to OBO Foundry¹ Principles, including:
 - Expressed in common syntax (OWL, XML)
 - Relations among ontology terms (*Is_A*, *Part_of*) are drawn from the OBO Relation ontology²
- **Granularity may be overly-fine for term annotation purposes:**
 - (Ogren, 2005) found that the terms which are used most often occur at higher levels of the ontology
 - **(Camon, et al. 2005) found that 35% of GO term names appeared as strings in a sample MEDLINE corpus.**

1 <http://www.obofoundry.org/crit.shtml>

2 <http://www.obofoundry.org/ro/>



MeSH Hierarchy: Annotation

- Annotation-related uses:
 - Annotating scientific papers from the PubMed/MEDLINE database. As documents are annotated with relevant concepts from the hierarchy, the documents themselves can be sorted, collected, organized, and retrieved more effectively.



MeSH Hierarchy: Granularity & Interoperability

- Interoperability is strong but weaker than GO:
 - Terms in the MeSH hierarchy are treated as topics in a partial ordering
 - **Parent concepts are “more specific than” child concepts, without constraints in terms of biological relations**
 - Leaf-level mappings are provided to comparable vocabularies, such as names in the Chemical Abstracts Service (CAS)¹
- Granularity is strongly linked to document-level annotation:

...MeSH contains a descriptor for 'Whales' but the domain of MeSH is biomedicine and not zoology. In the MEDLINE citation database, there are not sufficient citations to create a separate descriptor for each specific whale species. Nevertheless, it is useful to have the species names as entry terms to the descriptor. Gains in precision of retrieval by creating more specific descriptors would be small. (Nelson, 2001)
- **But may be too fine-grained for term-level annotation: depth > 11 in some sections**

1 <http://www.cas.org/expertise/cascontent/registry/>



GENIA Term Ontology: Annotation

- Annotation-related uses:
 - Biologically important terms are annotated in scientific abstracts that have been sampled from MEDLINE (latest release includes 18,545 annotated sentences)
 - Annotations are made as in-line XML markup
- Annotation example:

Covalent binding between N-acetyl-L-cysteine (NAC) and albumin was evaluated kinetically by conducting in vitro experiments.

Covalent <term sem="binding"> binding </term> between <term sem="Amino_acid_monomer"> N-acetyl-L-cysteine (NAC) </term> and <term sem="Protein_molecule"> albumin </term> was evaluated kinetically by conducting in vitro experiments.



GENIA Term Ontology: Granularity & Interoperability

- Interoperability is weaker than GO and MeSH:
 - Non-standard relations are used and applied inconsistently, so structural mapping to GO relations would be inexact
 - Term definitions are missing and/or unclear, so term mappings are difficult
 - Classification criteria among sibling classes are not consistent: sometimes biological, sometimes non-biological
 - Terms with common names have non-standard meanings (e.g. *Protein*)
- **Granularity is specifically tuned to terms that occur frequently in contiguous spans of text:**
 - Maximum depth of hierarchy is 6
 - Finer-grained distinctions would require more context (document-level) and possibly background knowledge
 - Candidate terms with few examples found in the course of annotation have been pruned



Plan of the Talk

- Background: Functional Requirements and KS
- GO, Mesh and GENIA
- Problems in GENIA
- New Scheme



GENIA Term Ontology: Granularity & Interoperability

- **Interoperability is weaker than GO (and MeSH):**
 - Non-standard relations are used and applied inconsistently, so structural mapping to GO relations would be inexact
 - Term definitions are missing and/or unclear, so term mappings are difficult
 - Classification criteria among sibling classes are not consistent: sometimes biological, sometimes non-biological
 - Terms with common names have non-standard meanings (e.g. *Protein*)
- Granularity is specifically tuned to terms that occur frequently in contiguous spans of text:
 - Maximum depth of hierarchy is 6
 - Finer-grained distinctions would require more context (document-level) and possibly background knowledge
 - Candidate terms with few examples found in the course of annotation have been pruned



Classifying Protein-related expressions

- **Different from classifying biological entities**
 - classification criteria are contextual, not biological
 - important point of confusion for users of GENIA
- Important for creating term-annotated data & training downstream systems
- **Expression classes** can be ambiguous in ways that biological entities are not, requiring common ancestors in the expression hierarchy:
 - **“NF Kappa B” Protein family? Protein complex?**
- Ontological definitions & principles can help distill these differences



Systematic Metonymy and Under-specification

Frequently observed ambiguities

- Gene Names and Gene Products (Proteins)
- Domain names and Proteins with the domains
- Family names and Specific protein molecules
- Protein Complexes and Pure Proteins

Under-specification in Text

- The local context (a sentence) often lacks information for disambiguation
- Some sentences refer to aggregated textual concepts which correspond to different biological entities:
 “This paper discusses A”, “the role of A in ...”



Classify *NF Kappa B* expressions with GENIA 0.9/1.0

Genia version 1.0 Term	Example sentence
<i>Protein_family_or_group</i>	“Fibrinogen activates NF-kappa B transcription factors in mononuclear phagocytes.”
<i>Protein_complex</i>	“Analysis of the nuclear extracts with antibodies directed against the major components of NF-kappa B , the p50 and RelA (p65) proteins, indicated that the composition of NF-kappa B was similar in neonatal and adult cells.”
<i>Individual protein molecule</i> (<i>Protein_molecule</i>)	“We have detected a specific nuclear protein complex that binds to the element and show that NF-kappa B1 (p50) is a part of this complex.”
<i>Subunit of protein complex</i>	<i>Deprecated in GENIA Term 1.0</i>
<i>Substructure of protein</i>	<i>Deprecated in GENIA Term 1.0</i>
<i>Protein_domain_or_region</i>	“Does nucleolin bind the NF kappa B DNA binding motif?”



Classify *NF Kappa B* expressions with GENIA 0.9/1.0

Genia version 1.0 Term	Example sentence
<i>Protein_family_or_group</i>	“Besides p50, 1,25(OH) ₂ D ₃ decreased the levels of another NF-kappa B protein , namely c-rel.”
<i>Protein_complex</i>	“This was due to the presence of active NF-kappa B in the nucleus of CD45- T cells.”
<i>Individual protein molecule</i> (<i>Protein_molecule</i>)	“In contrast, NF-kappa B (p50) alone fails to stimulate kappa B-directed transcription, and based on prior in vitro studies, is not directly regulated by I kappa B.”
<i>Subunit of protein complex</i>	<i>Deprecated in GENIA Term 1.0</i>
<i>Substructure of protein</i>	<i>Deprecated in GENIA Term 1.0</i>
<i>Protein_domain_or_region</i>	“Does nucleolin bind the NF kappa B DNA binding motif?”



Plan of the Talk

- Background: Functional Requirements and KS
- GO, Mesh and GENIA
- Problems in GENIA
- **New Scheme**



Principled ontology revision

We improve interoperability by:

- 1) Using standard biological definitions for terms in the biology taxonomy: for example “Protein”**
- 2) Using only biological `Is_A` in our relations, which helps us to do partial structural mapping between GENIA and other ontologies

We maintain expressiveness/granularity that we need for classifying text spans by:

- 1) Adding a new hierarchy where text-based classification features, like textual context and metonymic usage, can be explicitly represented**
- 2) Maintaining all of the classes that are used for GENIA annotation (although they have been re-arranged)



Expression types that support textual underspecification

- Add `As_collection` to the ontology, used to label expressions that use a protein name to refer to a collection
 - `As_collection`
 - `As_complex`
 - `As_family_or_group`
 - `As_substructure`
 - `As_domain_or_region`



Sentences that are hard to annotate show where underspecification is needed

Protein name with no expression type (molecule)	Induction of <term sem="Protein">NF-KB</term> during monocyte differentiation
Protein name used as a family name	...promotes induction or translocation of <term sem="Protein" type="As_family_or_group">NF-KB-related factors</term>
Protein name used as a chemical participant in an event	Dithiocarbamates (DTCs) have recently been reported as powerful inhibitors of <term sem="Protein" type="As_complex">NF-kappaB</term> activation ...
Ambiguous/difficult example: <i>NFAT</i> and <i>NF-KB</i> bear similar usage but different sem labels	... combined inhibitory effects on <term sem="Protein" type="As_family_or_group">NFAT</term> and <term sem="Protein" type="As_complex">NF-KB</term> support a potential use of DTCs...
NEW: <i>As_collection</i> indicates ambiguity between <i>As_complex</i> and <i>As_family_or_group</i>	... combined inhibitory effects on <term sem="Protein" type="As_collection">NFAT</term> and <term sem="Protein" type="As_collection">NF-KB</term> support a potential use of DTCs...



Why is this distinction so important?

- GENIA is often analyzed on the basis of its fitness to classify biological entities, which leads to confusion
- This distinction is part of the reason why GENIA definitions don't map to standard definitions of "Protein" (i.e. poor mapping to MeSH and GO)
- **Establishing which type of entities are annotated in GENIA (textual vs. biological) determines what tasks GENIA data can be used for: task definitions come from annotation principles**
- **Clarifying this issue helps annotators to understand their task (inter-annotator agreement is low for classes with an unclear textual/biological meaning!)**



Step 4: Add expression types that support textual underspecification

- Example strings where the annotation could be `As_complex` or `As_family_or_group`
- We know these expressions refer to either one or the other, but it's ambiguous which -> evidence that we need a new textual expression term

"ap-1"

"ap1"

"bcr"

"collagen"

"dimer"

"epo"

"erythropoietin"

"il-4r"

"mhc"

"nf-at"

"nf-kappab"

"nf-kappab/rel"

"nf-kb"

"globin"

"heterodimer"

"nf-y"

"nfat"

"pr"

"rel"

"rfx"

"tcr"

"histone"

"homodimer"



Evaluation

- Intrinsic criteria (current work):
 - Consistency of the term definitions with known biological classes: *Protein* is now consistent with common definition (and MeSH definition)
 - Consistent application of known biological relations: Is_A in the taxonomy is now OBO Is_A
- Empirical criteria (future work):
 - Quality of the annotations: inter-annotator agreement
 - Quality of the systems trained on the annotations: Define a task and see whether it supports the task
 - Example: named-entity recognition (Nédellec, 2006)



Notes & References

- Kim, Jin-Dong, Tomoko Ohta and Jun'ichi Tsujii. **Corpus annotation for mining biomedical events from literature**. BMC Bioinformatics. 9(1). pp. 10, BioMed Central, 2008. ISSN 1471-2105.
- Ogren, P. V., K. B. Cohen, et al. (2005). **Implications of Compositionality in the Gene Ontology for Its Curation and Usage**. Pacific Symposium on Biocomputing.
- Smith B, Ceusters W, Klagges B, Kohler J, Kumar A, Lomax J, Mungall CJ, Neuhaus F, Rector A, Rosse C **Relations in Biomedical Ontologies**. *Genome Biology*, 2005, 6:R46
- Camon, E. B., D. G. Barrell, et al. (2005). **An evaluation of GO annotation retrieval for BioCreAtIvE andGOA**. BMC Bioinformatics 6(Suppl 1)(S17).
- Nelson, S. J., D. Johnston, et al. (2001). **Relationships in Medical Subject Headings**. *Relationships in the organization of knowledge*. C. A. Bean and R. Green. New York, Kluwer Academic Publishers: 171-184.
- Nédellec, Claire, Philippe Bessières, Robert Bossy, Alain Kotoujansky, Alain-Pierre Manine (2006). **Annotation Guidelines for Machine Learning-Based Named Entity Recognition in Microbiology**. Proceedings of ECML/PKDD-2006 Workshop on Data and Text Mining for Integrative Biology.
- Schulz, S., E. Beißwanger, et al. (2006). **From GENIA to BioTop - Towards a top-level Ontology for Biology**. International Conference on Formal Ontology in Information Systems (FOIS 2006), Baltimore, USA.



Thank you.







Additional Topics (1)

Design a knowledge resource for text annotation (II)



Design a knowledge resource for text annotation (II)

- We may consider a more complex biological ontology, as long as we maintain consistent relation definitions and make the distinction between *expression types* and *biological types*
- Process:
 1. Start with a set of biological entities & relations that establish the domain of the knowledge resource
 2. Establish a cutoff level of granularity: biological entities that can be identified in referring expressions
 3. Find compositional properties of entities below the cutoff and summarize using new property ontologies
 4. Create a new ontology of *referring expression types* that links the biological classification with the textual classification
 5. Add new types to the expression ontology as needed, to support textual under-specification



Summarize compositional features in new ontology branches

- Find compositional properties of entities below the cutoff and summarize using new ontology terms
- Compositionality is a prominent feature of GO (Ogren, 2005), and we observe it in MeSH as well
- Example: Protein Families exist throughout MeSH, at varying depths
 - New top-level taxonomy term *Set_or_Collection* with *Protein_family* as a descendant
 - New relation *Has_member* relates a *Protein_family* to a *Protein*



Combine ontology terms to build complex annotations

<term sem="Protein_family">**TNF**</term>

<term sem="Protein_family">

 <term sem="Protein_substructure">**zinc
finger**</term> **proteins**</term>



Create a new hierarchy of expression types

- In this style, expression types are just functional definitions of retrieval tasks:
 - “*Protein referencing expressions*” will include instances of Protein, ProteinFamily and ProteinComplex.
 - “*Protein related expressions*” will include instances of Protein, ProteinFamily, ProteinComplex, proteinSubstructure and ProteinDomain, which corresponds to the Protein class in the current version of GENIA ontology.
- Genia_expression
 - Protein_referring_expression
 - Protein_related_expression



Additional Topics (2)

Compositionality in MeSH: Protein Families



Protein Families exist throughout MeSH

National Library of Medicine - Medical Subject Headings

2008 MeSH

[Return to Entry Page](#)

Please select a term from list:

[Armadillo Domain Proteins](#)

[Armadillo Protein Family](#)

[Chemokine CXCL11](#)

[Small Inducible Cytokine Subfamily B, Member 11 Protein](#)

[Ephrin-A2](#)

[Eph Ligand Family 1 Protein, Mammalian](#)

[Ephrin-B1](#)

[Eph Family Receptor Interacting Protein B1](#)

[Ephrin-B2](#)

[Eph Ligand Family 2 Protein, Mammalian](#)

[F-Box Proteins](#)

[F-Box Protein Family](#)

[Focal Adhesion Protein-Tyrosine Kinases](#)

[Focal Adhesion Protein-Tyrosine Kinase Family](#)

[Glial Cell Line-Derived Neurotrophic Factors](#)

[GDNF Protein Family](#)

[Glucose Transporter Type 1](#)

[Solute Carrier Family 2, Facilitated Glucose Transporter, Member 1 Protein](#)

[Glucose Transporter Type 2](#)

[Solute Carrier Family 2, Facilitated Glucose Transporter, Member 2 Protein](#)

[Glucose Transporter Type 3](#)

[Solute Carrier Family 2, Facilitated Glucose Transporter, Member 3 Protein](#)

[Glucose Transporter Type 4](#)

[Solute Carrier Family 2, Facilitated Glucose Transporter, Member 4 Protein](#)

[GTP-Binding Protein alpha Subunits, G12-G13](#)

[G-Protein, G12-G13 alpha Family](#)

[GTP-Binding Protein alpha Subunits, Gi-Go](#)

1

1

<http://www.nlm.nih.gov/mesh/2008/MBrowser.html>



Protein Families vary in depth

- Amino Acids, Peptides, and Proteins [D12]
 - Proteins [D12.776]

- Armadillo Domain Proteins [D12.776.091]
“Armadillo Protein Family”, “Arm Motif Proteins”
- Transcription Factors [D12.776.930]
 - Kruppel-Like Transcription Factors [D12.776.930.375]
 - Ikaros Transcription Factor [D12.776.930.375.500]
“IKAROS family zinc finger 1 protein, human”
 - Sp Transcription Factors [D12.776.930.375.750] +



Additional Topics (3)

Comparing Design Principles & Critiques of GENIA



Comparison: BioTop Design

A reasonable starting point for the ontological analysis of the biological upper-level is given by the following principles [5]: (i) select a set of foundational relations, (ii) define the ground axioms for these relations, (iii) establish constraints across the basic relations, (iv) define a set of formal properties induced by these formal relations, (v) introduce the basic categories and classify the relevant kinds of domain entities accordingly, and, finally, (vi) elicit the dependencies and interrelations among the basic categories. In our case, most of these basic categories are borrowed from the upper ontologies BFO [22] and DOLCE [6] enriched by principles introduced by Rector *et al.* [16]. Accordingly, we adopt the generally accepted, mutually exclusive divisions between universals and particulars on the one hand, and between continuants and occurrents on the other. Particulars (individuals) are the concrete and countable entities in the world (e.g., “my hand”) whereas universals are entities which are instantiated by particulars (e.g., “hand”²).



BioTop (OBO Foundry) Relations

- *Instance_of*: <instance, class>
- *Is-A*: <class, class>
- *Part_of*: <instance, instance>
- *Has_part*: <instance, instance>
- *Derives_from*: <instance, instance>
- *Has_function*: <class, function>
- *Inheres*: <function, class>



Additional Topics (4)

Setting a granularity threshold for the new ontology



The resulting sub-tree for Transcription Factors, Cells, & Humans

- Chemicals and Drugs [D]
 - Amino acids, peptides, and proteins (and ~16 siblings)
 - proteins (and 2 siblings)
 - transcription factors (and ~93 siblings)
 - NF-kappa B (and ~45 siblings)
- Anatomy [A]
 - Cells [A11] (and ~17 siblings)
- Organisms [B]
 - Animals [B01]
 - Chordata [B01.150] (and 2 siblings)
 - Vertebrates [B01.150.900] (and 1 sibling)
 - Mammals [B01.150.900.649](and 4 siblings)
 - Primates (and 15 siblings)
 - Haplorhini
 - Catarrhini
 - Hominidae
 - Humans (and 4 siblings)



Extra Slides (1)

Annotated examples of “NF-KappaB” from the GENIA corpus



Full-sentence examples of “NF-KB” from GENIA_pos_term_40a/* .xml

HIV1 infection of <term sem="Cell_natural" id="T1" lex="human_monocyte">human monocytes</term> and<term sem="Cell_natural" id="T2" lex="macrophage">macrophages</term> promotes induction or translocation of <term sem="Protein_family_or_group" id="T3" lex="NF-KB-related_factor">NF-KB-related factors</term>.

GENIA_pos_term_40a/1984449.xml:9:<ArticleTitle>Induction of <term sem="Protein_molecule" id="T1" lex="NF-KB">NF-KB</term> during <term id="T2" lex="monocyte_differentiation"><term sem="Cell_natural" id="T3" lex="monocyte">monocyte</term> differentiation</term> by <term id="T4" lex="HIV_type_1_infection"><term sem="Virus" id="T5" lex="HIV_type_1">HIV type 1</term> infection</term>.

Dithiocarbamates</term> (<term sem="Organic_compound_other" id="T5" lex="DTC">DTCs</term>) have recently been reported as powerful inhibitors of <term id="T6" lex="NF-kappaB_activation"><term sem="Protein_complex" id="T7" lex="NF-kappaB">NF-kappaB</term> activation</term> in a number of cell types.



Full-sentence examples of “NF-KB” from GENIA_pos_term_40a/* .xml (2)

Given the role of this transcription factor in the regulation of <term id="T8" lex="gene_expression">gene expression</term> in the <term id="T9" lex="inflammatory_response">inflammatory response</term>, <term sem="Organic_compound_other" id="T10" lex="NF-kappaB_inhibitor"><term sem="Protein_complex" id="T11" lex="NF-kappaB">NF-kappaB</term> inhibitors </term> have been suggested as potential <term sem="Organic_compound_other" id="T12" lex="therapeutic_drug">therapeutic drugs</term> for <term id="T13" lex="inflammatory_disease">inflammatory diseases</term>

... and involved not only the inhibition of <term id="T26" lex="NF-kappaB-driven_reporter_activation"><term sem="Protein_complex" id="T27" lex="NF-kappaB">NF-kappaB</term>-driven reporter activation</term> but also that of

In addition, the combined inhibitory effects on <term sem="Protein_family_or_group" id="T62" lex="NFAT">NFAT</term> and <term sem="Protein_complex" id="T63" lex="NF-KB">NF-KB</term> support a potential use of <term sem="Organic_compound_other" id="T64">



Extra Slides (2)

Discussion issues with GENIA 0.9/1.0



Resolve conflicts with ontological principles

- The is-a relation is violated
- The subclasses are not exhaustive
- Relationships btw the subclasses are not defined at all (not clear to users or to designers)
- Relationships among “family-or-group” subclasses among DNA, RNA, and Protein are unclear (although something like this exists in GO, as per the compositionality paper)



Ad-hoc protein annotation

- Annotate mentions of proteins and protein-related concepts
 - Example sentence tagged with Protein & with a few subclasses
 - When subclasses can be distinguished, propose new ones as needed
 - When subclasses lose coherence, merge their instances into parent or sibling classes



Extra Slides (3)

Gene Ontology



Granularity of Gene Ontology

- Terms are specific enough to support domain modeling on a fine scale

all : all [250226 gene products]

⊕ ⓘ GO:0003674 : molecular_function [168625 gene products]

⊕ ⓘ GO:0003824 : catalytic activity [52022 gene products]

⊕ ⓘ GO:0016740 : transferase activity [15746 gene products]

⊕ ⓘ GO:0016772 : transferase activity, transferring phosphorus-containing groups [7975 gene products]

⊕ ⓘ GO:0016301 : kinase activity [6091 gene products]

⊕ ⓘ GO:0004672 : protein kinase activity [3505 gene products]

⊕ ⓘ GO:0004713 : protein tyrosine kinase activity [402 gene products]

⊕ ⓘ **GO:0004716 : receptor signaling protein tyrosine kinase activity [21 gene products]**

- Terms may represent entities or events that are only implicitly present in any single span of text
- Total size > 10,000 terms arranged in a hierarchy of depth [0,7]



Interoperability of the Gene Ontology

- Conforms to OBO principles¹, including the following examples:
 - Open, license-free access
 - Expressed in common shared syntax (OWL, XML, etc.)
 - Clearly delineated content domain (non-overlapping with other OBO ontologies)
 - Relations among ontology terms (*Is_A*, *Part_of*) are drawn from the OBO Relation ontology²
- Open annotation/development model contributes to interoperability

1 <http://www.obofoundry.org/crit.shtml>

2 (Smith, et al. 2005)



Extra Slides (4)

MeSH Hierarchy



Extra Slides (5)

Revising the GENIA Term Ontology



Design a knowledge resource for annotating & classifying expressions

- Unify the demands of biological / textual classification
- Leverage ontological design principles to make the resource more usable
 - Establish consistent relation definitions that conform to ontological practices
 - Entity definitions that draw on existing biological ontologies
- Process:
 1. Start with a taxonomy of biological entities that establish the domain of the knowledge resource
 2. Establish a cutoff level of granularity: biological entities that can be identified in referring expressions
 3. Create a new ontology of *referring expression types* that links the biological classification with the textual classification
 4. Add new types to the expression ontology as needed, to support textual under-specification

