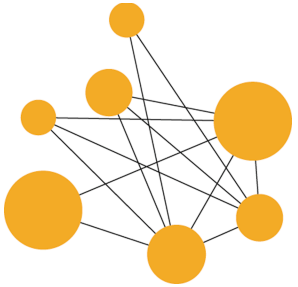
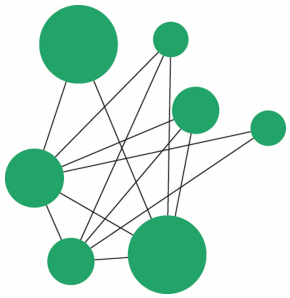
Ontology Quality Assurance via Term Transformations

Karin Verspoor, Daniel Dvorkin,
K. Bretonnel Cohen, and Lawrence Hunter

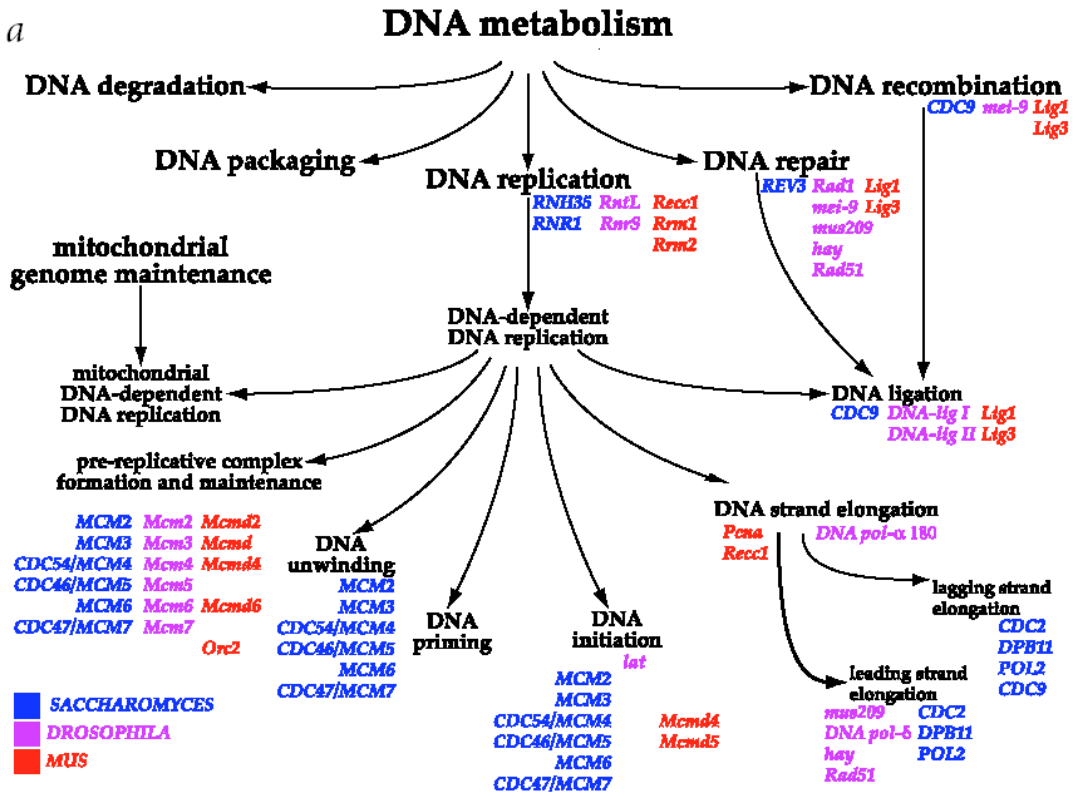




Karin.Verspoor@ucdenver.edu
http://compbio.uchsc.edu/Verspoor

Community-curated Ontologies

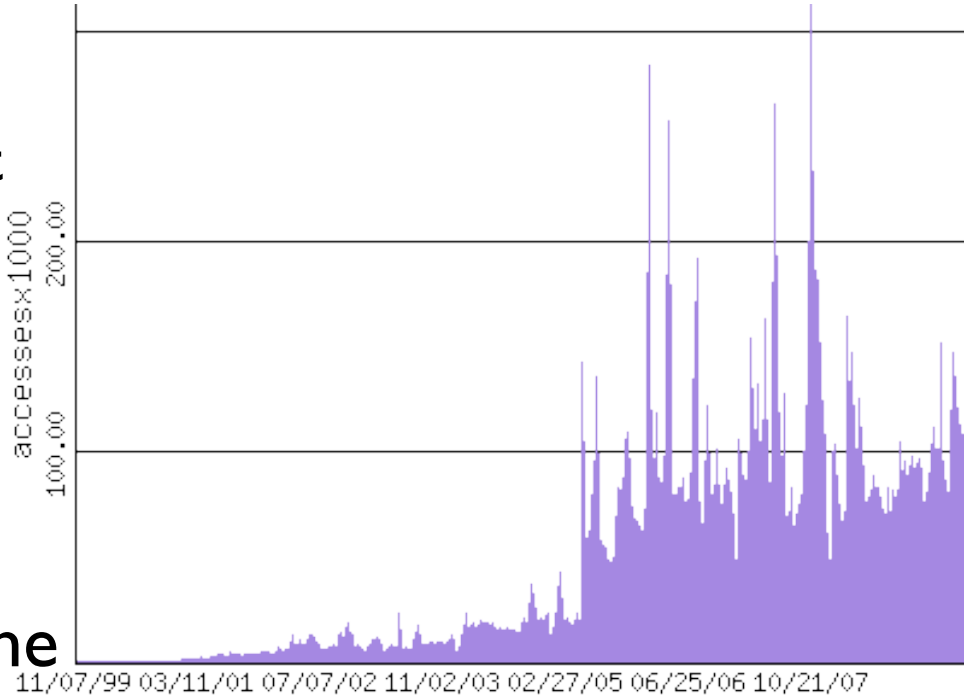
- Controlled vocabularies, with hierarchical backbone
- Used to facilitate consistent description in databases across organisms



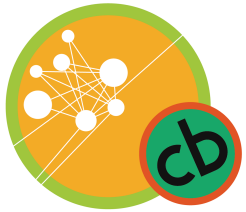
Portion of the Gene Ontology

The Gene Ontology: Usage

- 27602 terms
  - 16644 biological_process
  - 2359 cellular_component
  - 8599 molecular_function
- Gene Annotations for 40+ organisms
- 2068 publications in PubMed matching "gene ontology"
- ISI Web of Knowledge: 3781 refs to GO paper



Statistics as of June 9, 2009



Key quality concern: *Univocality*

- Univocality = *one voice* (Spinoza, 1677)

  "a shared interpretation of the nature of reality"

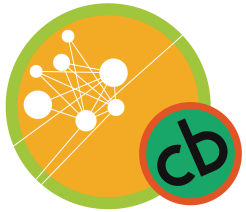
  (with thanks to David Hill @ Jackson Lab)
- Consistency of expression of concepts
- Regular, compositional, linguistic structure
  - Facilitates human usability
  - Computational tools can utilize this regularity

Regulation of transcription

Transcription Regulation

Positive regulation of cell migration

Cell migration positive regulation



Quality Assurance in the GO

- Goal: identify violations of univocality
- Problem: the GO is generally very high quality; how to identify the few inconsistencies?
- Hypothesis: violations of univocality will correspond to transformational variants
- Strategy: term transformation & clustering

GO Term Transformation: Abstraction

- Substitution of embedded GO & ChEBI terms

toluene oxidation via 3-hydroxytoluene

CTERM oxidation via CTERM

regulation of coagulation

regulation of GTERM

leukotriene production during acute inflammatory response

CTERM production during GTERM

GO Term Transformations

- Stopword removal

  toluene oxidation via 3-hydroxytoluene

  toluene oxidation 3-hydroxytoluene

  regulation of coagulation

  regulation coagulation

- Alphabetic reording

  3-hydroxytoluene oxidation toluene via

  coagulation of regulation

Transformation combinations

- Abstraction=1, StopRemoval=1, Reordering=1

toluene oxidation via 3-hydroxytoluene

regulation of coagulation

leukotriene production during acute inflammatory response

Transformation combinations

- Abstraction=1, StopRemoval=1, Reordering=1

toluene oxidation ~~via~~ 3-hydroxytoluene

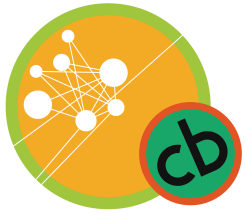
CTERM CTERM oxidation

regulation ~~of~~ coagulation

GTERM regulation

leukotriene production ~~during~~ acute inflammatory response

CTERM GTERM production

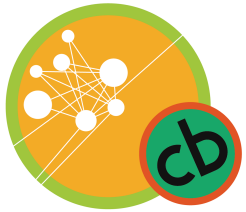


Clustering

- Group together all terms with a common form *after transformation*
- Perform clustering for different combinations of transformations

asr {GTERM constit structu}

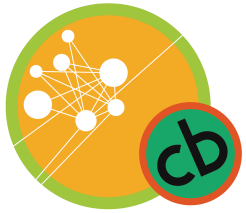
GO:0005201 -- extracellular matrix structural constituent
GO:0005199 -- structural constituent of cell wall
GO:0005213 -- structural constituent of chorion
GO:0005200 -- structural constituent of cytoskeleton
GO:0003735 -- structural constituent of ribosome
GO:0017056 -- structural constituent of nuclear pore
GO:0019911 -- structural constituent of myelin sheath



Analysis of clusters

	num. clusters	proportion
Total candidates	237	
Identical	47	
False Positive (FP)	123	65%
True Positive (TP)	67	35%

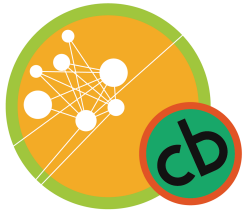
- Heuristic search:
  - Consider only clusters with abstraction (a±±)
  - Identify terms in distinct a-- clusters, but merge together in a-r, as-, or asr.
- Manual assessment of 190 clusters



Transformation Impact

xyz	count	mean	max	xyz	count	mean	max
—	23,478	1.088	29	a–	12,704	2.010	2999
–r	23,395	1.092	29	a-r	12,594	2.028	3003
-s-	23,400	1.091	31	as-	12,564	2.033	3012
-sr	23,294	1.096	31	asr	12,354	2.067	3054

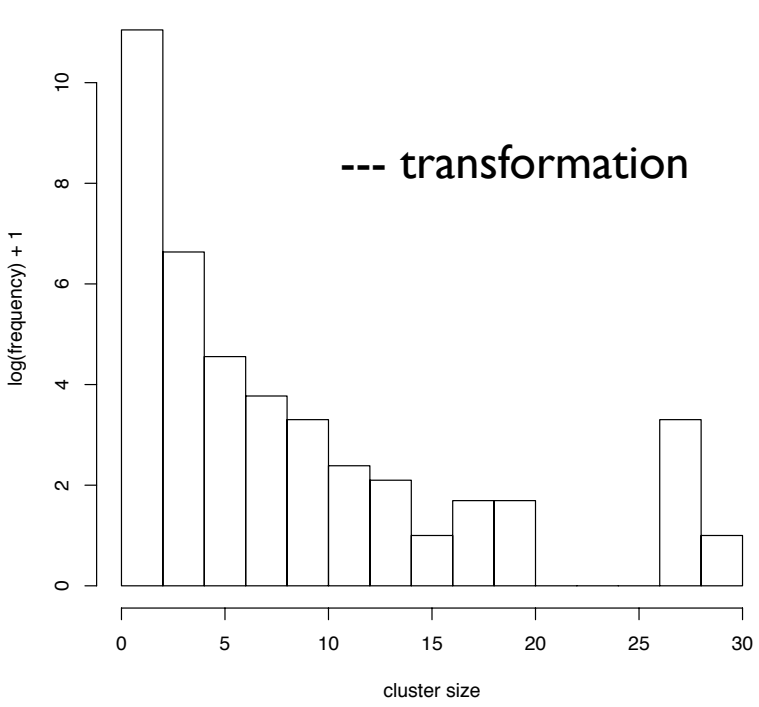
- 25,539 source GO terms (12/2007 version)
- Pre-processing reduces to 23,478 (8%)
- a=Abstraction, s=StopRemoval, r=Reordering
- Abstraction has most impact: 46% reduction

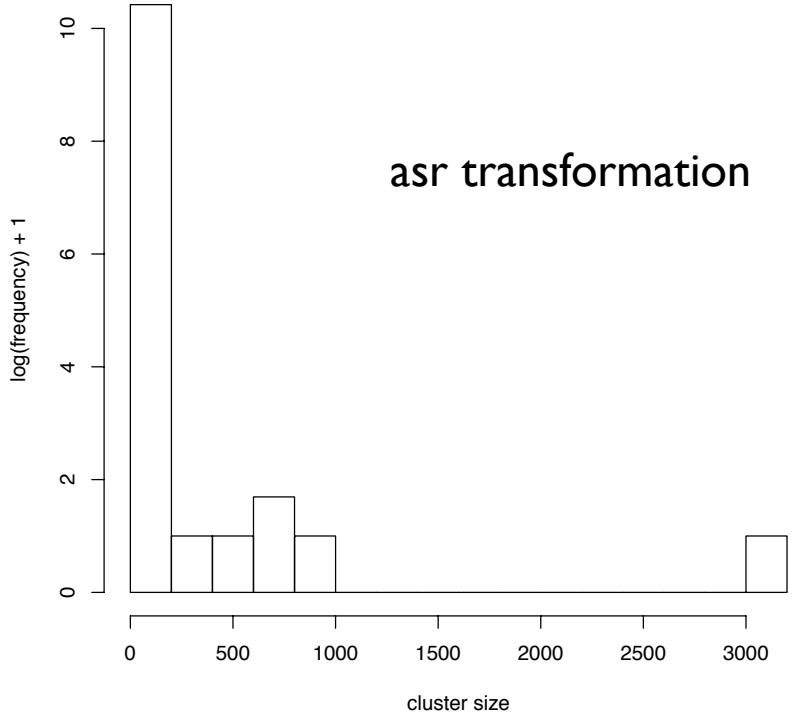


Abstraction breakdown, a-- clusters

abstraction	count	percentage
CTERM only	2,489	20%
GTERM only	3,840	30%
Both CTERM & GTERM	1,415	11%
no abstraction	4,960	39%

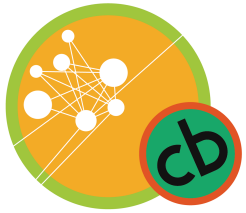
Distribution of cluster size





True Positive clusters

- 67 clusters
- 317 GO terms
- Obsolete term filter: 7 clusters, 32 terms
- Approximately 77 term rephrasings anticipated



True Positive inconsistencies

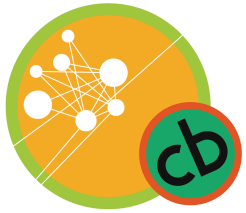
- {X Y} ≈ {Y of X} | {Y in X} [45%]

{GTERM GTERM organis symbion}

GO:0052387 -- induction by organism of symbiont apoptosis
GO:0052351 -- induction by organism of systemic acquired resistance in symbiont
GO:0052350 -- induction by organism of induced systemic resistance in symbiont
GO:0052560 -- induction by organism of symbiont immune response
GO:0052399 -- induction by organism of symbiont programmed cell death
GO:0052396 -- induction by organism of symbiont non-apoptotic programmed cell death

{GTERM multice organis}

GO:0010259 -- multicellular organismal aging
GO:0022412 -- reproductive cellular process in multicellular organism
GO:0032504 -- multicellular organism reproduction
GO:0033057 -- reproductive behavior in a multicellular organism
GO:0033555 -- multicellular organismal response to stress
GO:0035264 -- multicellular organism growth



True Positives (2)

- Determiners [16%]

{GTERM forebra}

GO:0021861 -- radial glial cell differentiation in the forebrain

GO:0021846 -- cell proliferation in forebrain

GO:0021872 -- generation of neurons in the forebrain

{GTERM organ}

GO:0031100 -- organ regeneration

GO:0035265 -- organ growth

GO:0010260 -- organ senescence

GO:0001759 -- induction of an organ

True Positives (3)

- Other alternations [16%]

{GTERM selecti site}

GO:0000282 -- cellular bud site selection

GO:0000918 -- selection of site for barrier septum formation

- Conflicting conventions [6%]

{GTERM endothe} (partial listing)

GO:0003100 -- regulation of systemic arterial blood pressure by endothelin

GO:0004962 -- endothelin receptor activity

- Punctuation [3%]

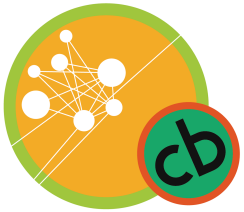
GO:0016653 -- oxidoreductase activity, acting on NADH, heme protein as acceptor

GO:0016658 -- oxidoreductase activity, acting on NADH, flavin as acceptor

GO:0050664 -- oxidoreductase activity, acting on NADH, with oxygen as acceptor

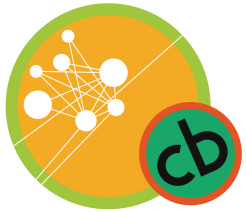
GO:0043247 -- telomere maintenance in response to DNA damage

GO:0042770 -- DNA damage response, signal transduction



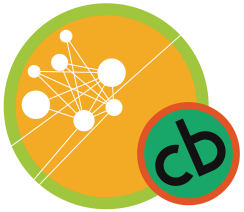
True Positives (4)

- "Grab bag"
  - Lexical choice
    - "within" vs. "in"
    - "substrate-specific" vs. "substrate-dependent"
  - Superfluous words like "other"



False positive breakdown

	num. clusters	FP proportion
semantic import of stopword	61	50%
non-parallel structure	33	27%
semantic import of stemming	21	17%
syntactic variation	6	5%
semantic import of word order	1	1%
mis-classified content word	1	1%



False positive cluster examples

- Semantic import of stopword [50%]

{CTERM GTERM levels modulat symbion} (partial listing)

GO:0052430 – modulation by host of symbiont RNA levels

GO:0052018 – modulation by symbiont of host RNA levels

{CTERM CTERM galacto GTERM}

GO:0033580 -- protein amino acid galactosylation at cell surface

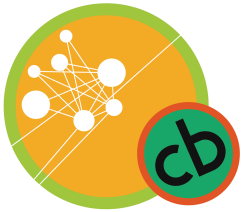
GO:0033582 -- protein amino acid galactosylation in cytosol

GO:0033579 -- protein amino acid galactosylation in endoplasmic reticulum

{callose deposit GTERM}

GO:0052542 -- callose deposition during defense response

GO:0052543 -- callose deposition in cell wall



False positives (2)

- Non-parallel structure [27%]

{CTERM CTERM}

GO:0005204 -- chondroitin sulfate proteoglycan

GO:0006088 -- acetate to acetyl-CoA

GO:0015641 -- lipoprotein toxin

{GTERM GTERM GTERM} (partial listing)

GO:0019896 -- axon transport of mitochondrion

GO:0047496 -- vesicle transport along microtubule

GO:0047497 -- mitochondrion transport along microtubule

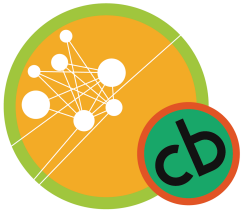
GO:0032066 -- nucleolus to nucleoplasm transport

GO:0052067 -- negative regulation by symbiont of entry into host cell via phagocytosis

{GTERM storage}

GO:0001506 -- neurotransmitter biosynthetic process and storage

GO:0000322 -- storage vacuole



False positives (3)

- Stemming [17%]

{regulat GTERM} (partial listing)

GO:0045066 -- regulatory T cell differentiation

GO:0045069 -- regulation of viral genome replication

GO:0045055 -- regulated secretory pathway

GO:0031347 -- regulation of defense response

- Syntactic variation [5%]

{GTERM mainten}

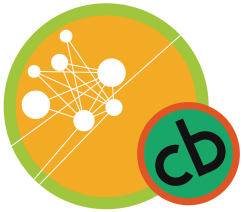
GO:0045216 -- intercellular junction assembly and maintenance

GO:0045217 -- intercellular junction maintenance

GO:0045218 -- zonula adherens maintenance

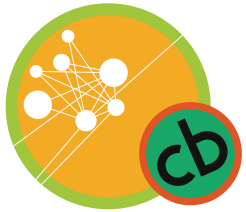
- Semantic import of word order[5%]

{GTERM CTERM activit}	{CTERM GTERM activit}
apoptosis inhibitor activity	gibberellin binding activity



Conclusions

- Used simple term transformations and heuristic search
- Able to reduce set of clusters to be manually evaluated to 190 (for 25k terms)
- Identified 67 TP instances of univocality violations covering 317 GO terms
- Future work
  - More specific linguistic alternations
  - Improve heuristics for TP search



Acknowledgements

- Mike Bada (Ontologist, CRAFT)
- Bill Baumgartner (Research engineer)
- Lynne Fox (Librarian)
- Helen Johnson (Linguist)
- Chris Roeder (Software engineer)
- Hannah Tipney (Analyst)

- NIH grants
  - R01 GM 083649
  - T15 LM 009451

Opportunities at one of the best Computational Bioscience Programs

- Top faculty, great research, serious education
- Institutional Training Grant from NLM
- Two open postdocs to be filled this summer
- Grad school application deadline January 1
- Open faculty positions
- More info at http://compbio.uchsc.edu
- Ask me or Larry Hunter for details