

New York State Center of Excellence
in **Bioinformatics**
& Life Sciences



UB **University at Buffalo**
The State University of New York
College of Arts & Sciences

Workshop on bio-ontologies
October 28, 2005

Enriching Bio-ontologies with Non-hierarchical Relations



Olivier Bodenreider

Lister Hill National Center
for Biomedical Communications
Bethesda, Maryland - USA

Acknowledgments



◆ Marc Aubry
*UMR 6061 CNRS,
Rennes, France*



◆ Anita Burgun
*University of
Rennes, France*



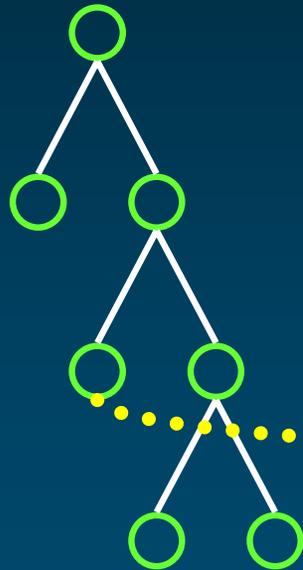
Gene Ontology

- ◆ Annotate gene products
- ◆ Coverage
 - Molecular functions
 - Cellular components
 - Biological processes
- ◆ Explicit relations to other terms within the same hierarchy
- ◆ No (explicit) relations
 - To terms across hierarchies
 - To concepts from other biological ontologies

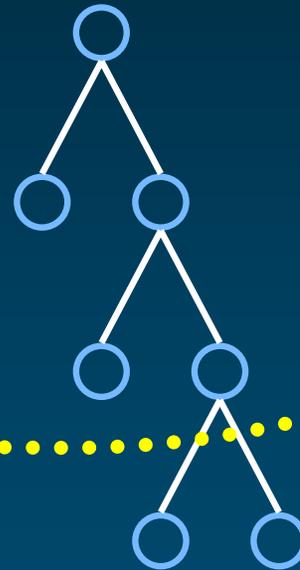


Gene Ontology

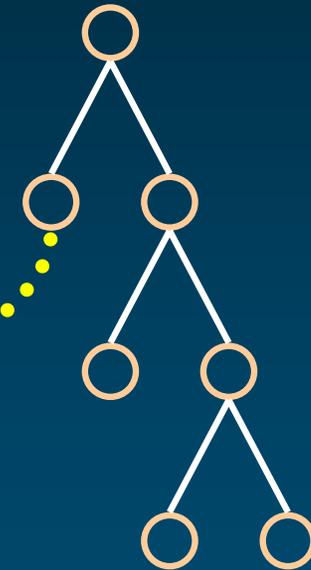
Molecular
functions



Cellular
components



Biological
processes



BP: metal ion transport
MF: metal ion transporter activity



Motivation

- ◆ Richer ontology
 - Beyond hierarchies
- ◆ Easier to maintain
 - Explicit dependence relations
- ◆ More consistent annotations
 - Quality assurance
 - Assisted curation



Related work

◆ Ontologizing GO

- GONG [Wroe & al., PSB 2003]

◆ Identifying relations among GO terms across hierarchies

- Lexical approach [Ogren & al., PSB 2004-2005]
- Non-lexical approaches [Bodenreider & al., PSB 2005]

◆ Identifying relations between GO terms and OBO terms

- ChEBI [Burgun & al., SMBM 2005]

◆ Representing relations among GO terms and between GO terms and OBO terms

- Obol [Mungall, CFG 2005]

◆ See also: [Bada & al., 2004], [Kumar & al., 2004], [Dolan & al., 2005]



Non-lexical approaches to identifying associative relations in the Gene Ontology

GO and annotation databases

◆ 5 model organisms

- FlyBase
- GOA-Human
- MGI
- SGD
- WormBase



270,000 gene-term associations

Brca1	GO:0000793	IDA
Brca1	GO:0003677	IEA
Brca1	GO:0003684	IDA
Brca1	GO:0003723	ISS
Brca1	GO:0004553	ISS
Brca1	GO:0005515	ISS, TAS
Brca1	GO:0005622	IEA
Brca1	GO:0005634	IDA, ISS
Brca1	...	



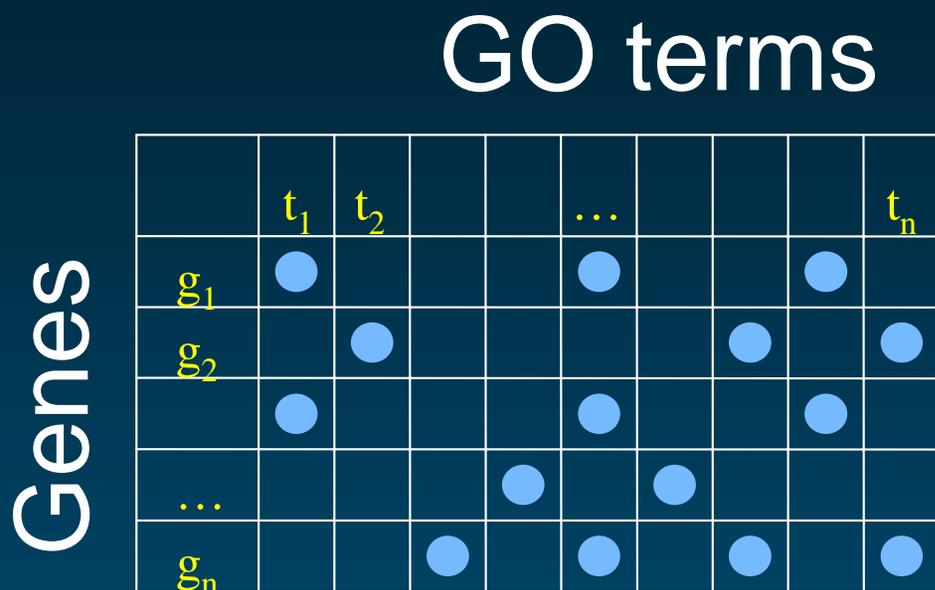
Three non-lexical approaches

All based on annotation databases

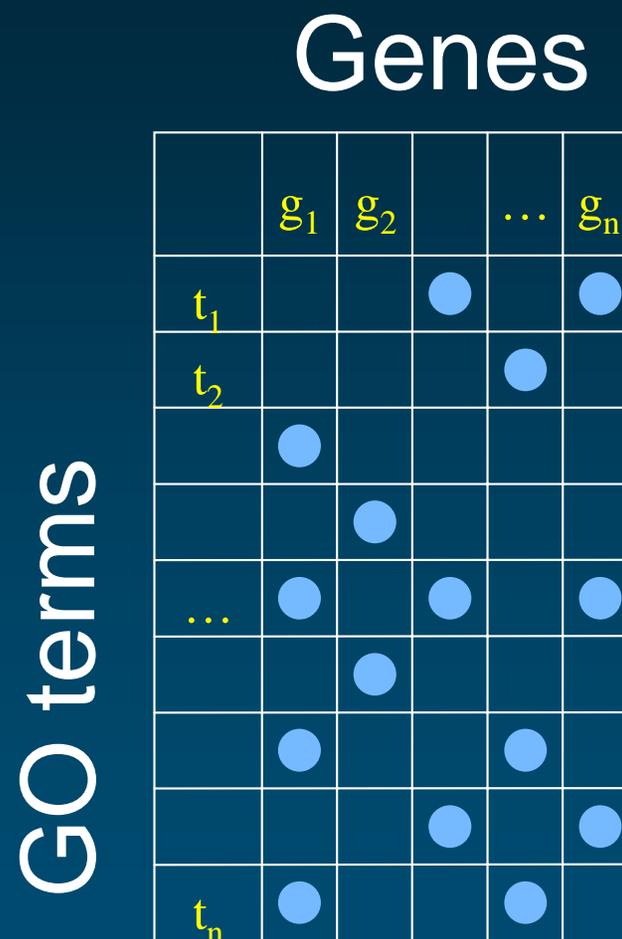
- ① Similarity in the vector space model
- ② Statistical analysis of co-occurring GO terms
- ③ Association rule mining



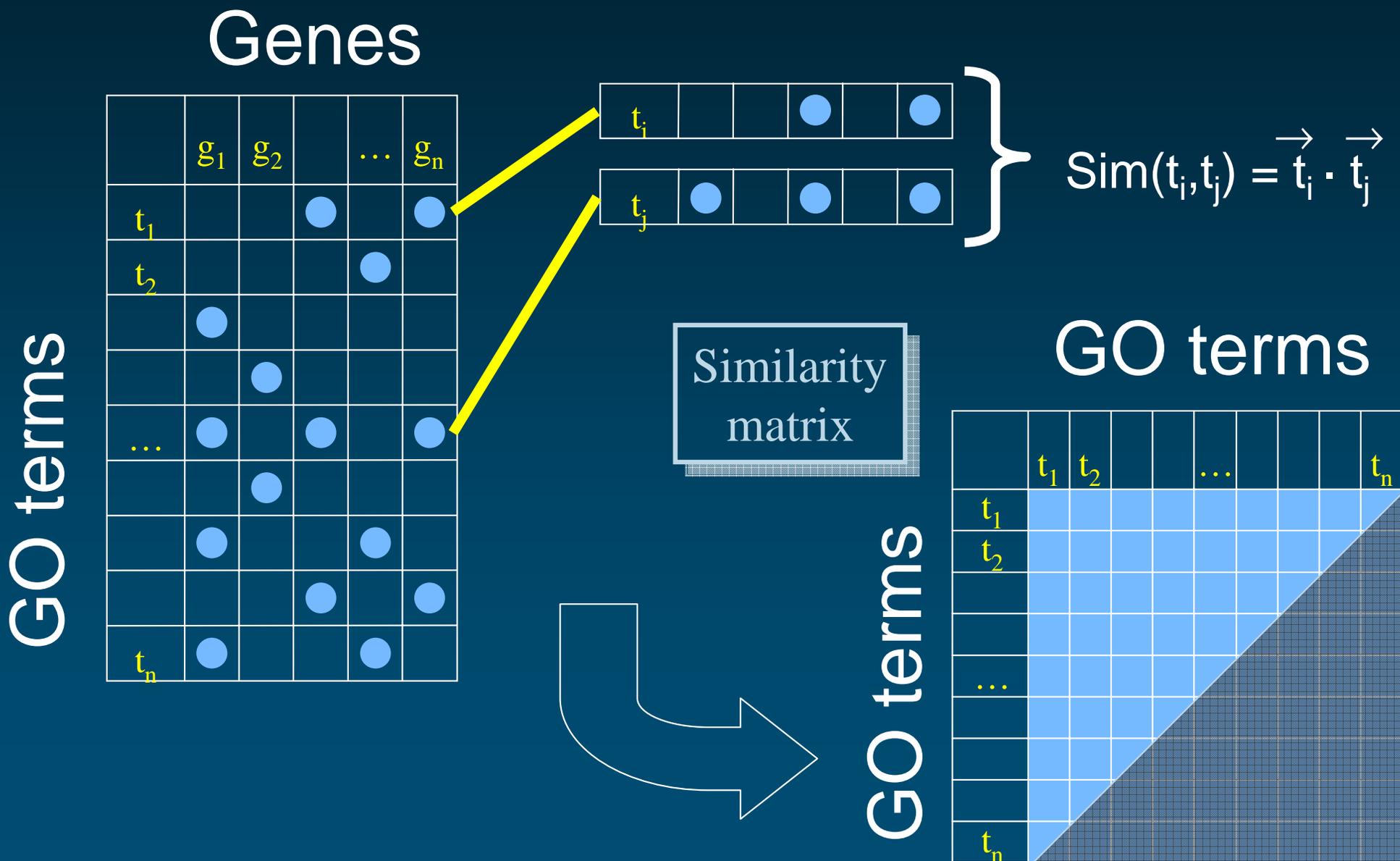
1 Similarity in the vector space model



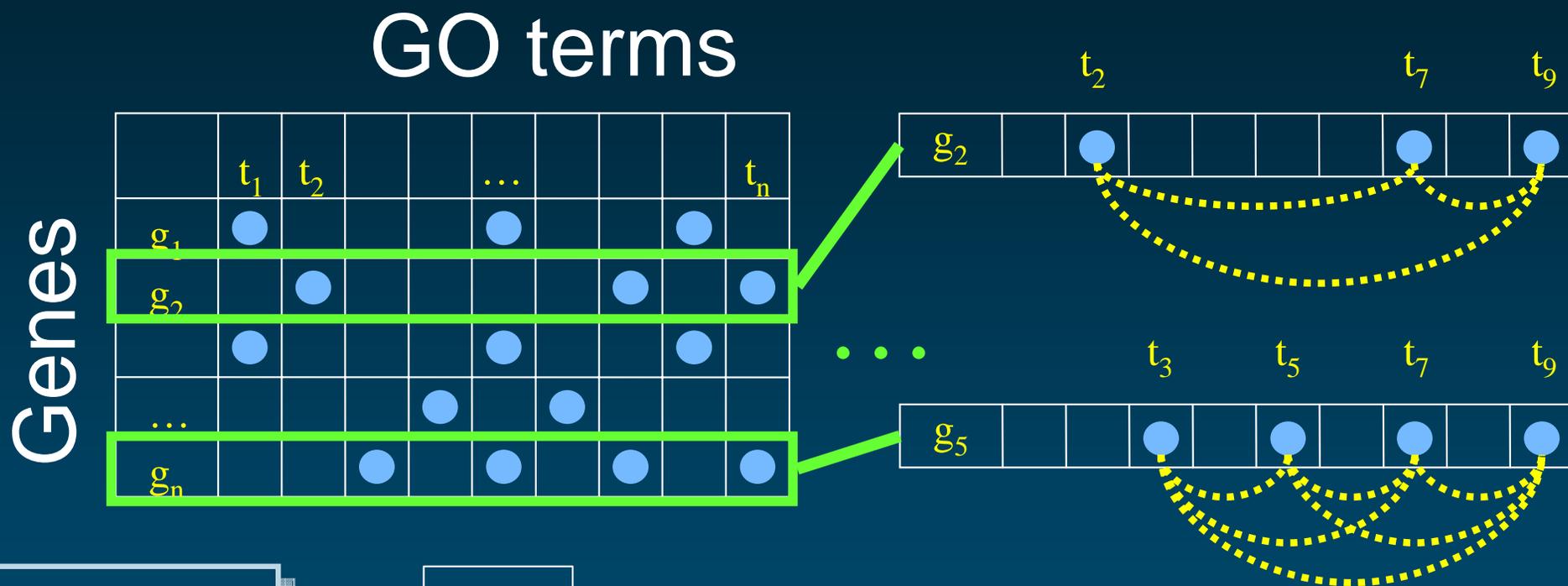
Annotation
database



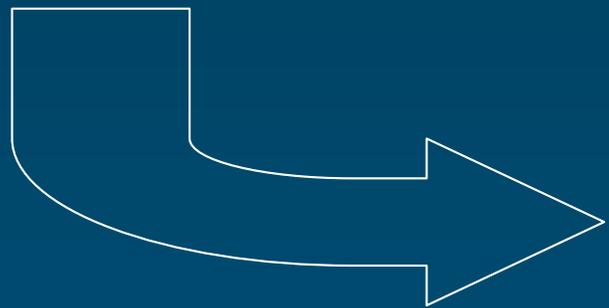
1 Similarity in the vector space model



2 Analysis of co-occurring GO terms



Annotation database



t ₂ -t ₇	1
t ₂ -t ₉	1
t ₇ -t ₉	2
...	

t ₅	1
t ₇	2
t ₉	2
...	

2 Analysis of co-occurring GO terms

◆ Statistical analysis: test independence

- Likelihood ratio test (G^2)
- Chi-square test (Pearson's χ^2)

◆ Example from GOA (22,720 annotations)

- C0006955 [BP] Freq. = 588
 - C0008009 [MF] Freq. = 53
- } Co-oc. = 46

GO:0008009 *immune response*

	present	absent	Total
GO:0006955 <i>chemokine activity</i>	46	542	588
	7	21,583	22,132
	53	22,125	22,720

$$G^2 = 298.7$$
$$p < 0.000$$

3

Association rule mining

GO terms

Genes

	t_1	t_2			...			t_n
g_1	●				●			●
g_2		●					●	●
	●				●			●
...				●		●		
g_n			●		●		●	●



transaction

Annotation database



apriori

- Rules: $t_1 \Rightarrow t_2$
- Confidence: $> .9$
- Support: $.05$

Examples of associations

Association		VSM	COC	ARM	LEX
MF: <i>potassium channel activity</i>	[GO:0005267]	X	X	X	
BP: <i>potassium ion transport</i>	[GO:0006813]				
MF: <i>chemokine activity</i>	[GO:0008009]		X	X	
BP: <i>immune response</i>	[GO:0006955]				
CC: <i>hemoglobin complex</i>	[GO:0005833]	X	X		
BP: <i>oxygen transport</i>	[GO:0015671]				
MF: <i>taste receptor activity</i>	[GO:0008527]	X		X	
BP: <i>perception of taste</i>	[GO:0050909]				
MF: <i>metal ion transporter activity</i>	[GO:0046873]	X		X	X
BP: <i>metal ion transport</i>	[GO:0030001]				
CC: <i>transport vesicle</i>	[GO:0030133]				X
BP: <i>transport</i>	[GO:0006810]				
CC: <i>gap junction</i>	[GO:0005921]	X	X		
BP: <i>cell communication</i>	[GO:0007154]				

Associations identified

	VSM	COC	ARM	LEX
MF-CC	499	893	362	917
MF-BP	3057	1628	577	2523
CC-BP	760	1047	329	2053
Total	4316	3568	1268	5493

7665 by at least one approach



Associations identified VSM

	VSM	COC	ARM	LEX
MF-CC	499	893	362	917
MF-BP	3057	1628	577	2523
CC-BP	760	1047	329	2053
Total	4316	3568	1268	5493

MF: ice binding

BP: response to freezing



Associations identified COC

	VSM	COC	ARM	LEX
MF-CC	499	893	362	917
MF-BP	3057	1628	577	2523
CC-BP	760	1047	329	2053
Total	4316	3568	1268	5493

MF: chromatin binding
CC: nuclear chromatin



Associations identified ARM

	VSM	COC	ARM	LEX
MF-CC	499	893	362	917
MF-BP	3057	1628	577	2523
CC-BP	760	1047	329	2053
Total	4316	3568	1268	5493

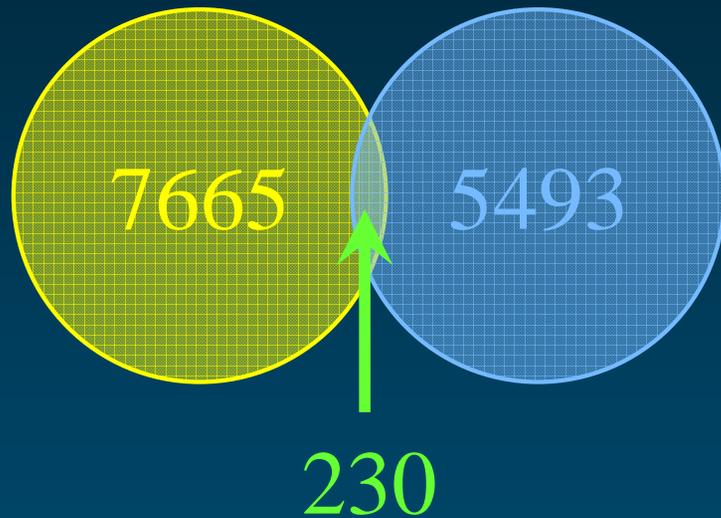
MF: carboxypeptidase A activity

BP: peptolysis and peptidolysis

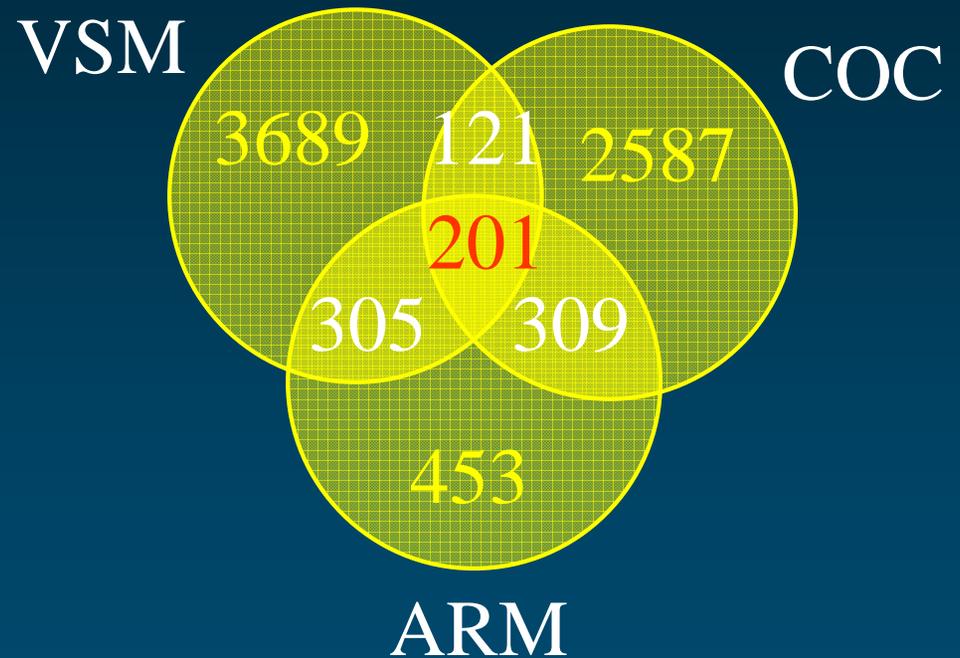


Limited overlap among approaches

◆ Lexical vs. non-lexical



◆ Among non-lexical



Linking the Gene Ontology to other biological ontologies

Related domains

- ◆ Organisms cytosolic ribosome (sensu Eukaryota)
- ◆ Cell types T-cell activation
- ◆ Physical entities
 - Gross anatomy brain development
 - Molecules transferrin receptor activity
- ◆ Functions
 - Organism functions visual perception
 - Cell functions T-cell activation
- ◆ Pathology regulation of blood pressure



GO and other domains

	Physical entity	Function	Process
Organism	Gross anatomy	Organism functions	Organism processes
Cell	Cellular components	Cellular functions	Cellular processes
Molecule	Molecules	Molecular functions	Molecular processes

[adapted from B. Smith]



GO and other domains (revisited)

Resolution	Physical whole	Physical part	Function	Process
	Organism	Organism components	Organism functions	Organism processes
	Cell	Cellular components	Cellular functions	Cellular processes
	Molecule	Molecular components	Molecular functions	Molecular processes

[adapted from B. Smith]



Biological ontologies (OBO)

Domain	Prefix	Files
Cell type	CL	cell.obo
Chemical entities of biological interest	CHEBI	ontology.obo
Mus adult gross anatomy	MA	MA.ontology
Plant anatomy	PO	anatomy.ontology and anatomy.definition
NCBI organismal classification	taxon	taxonomy.dat
Human disease	DOID	DO_08_18_03.txt
Mouse pathology	MPATH	mouse_pathology.ontology
PATO	PATO	attribute_and_value.obo
Physical-chemical methods and properties	FIX	fix.ontology
Physico-chemical process	REX	rex.obo

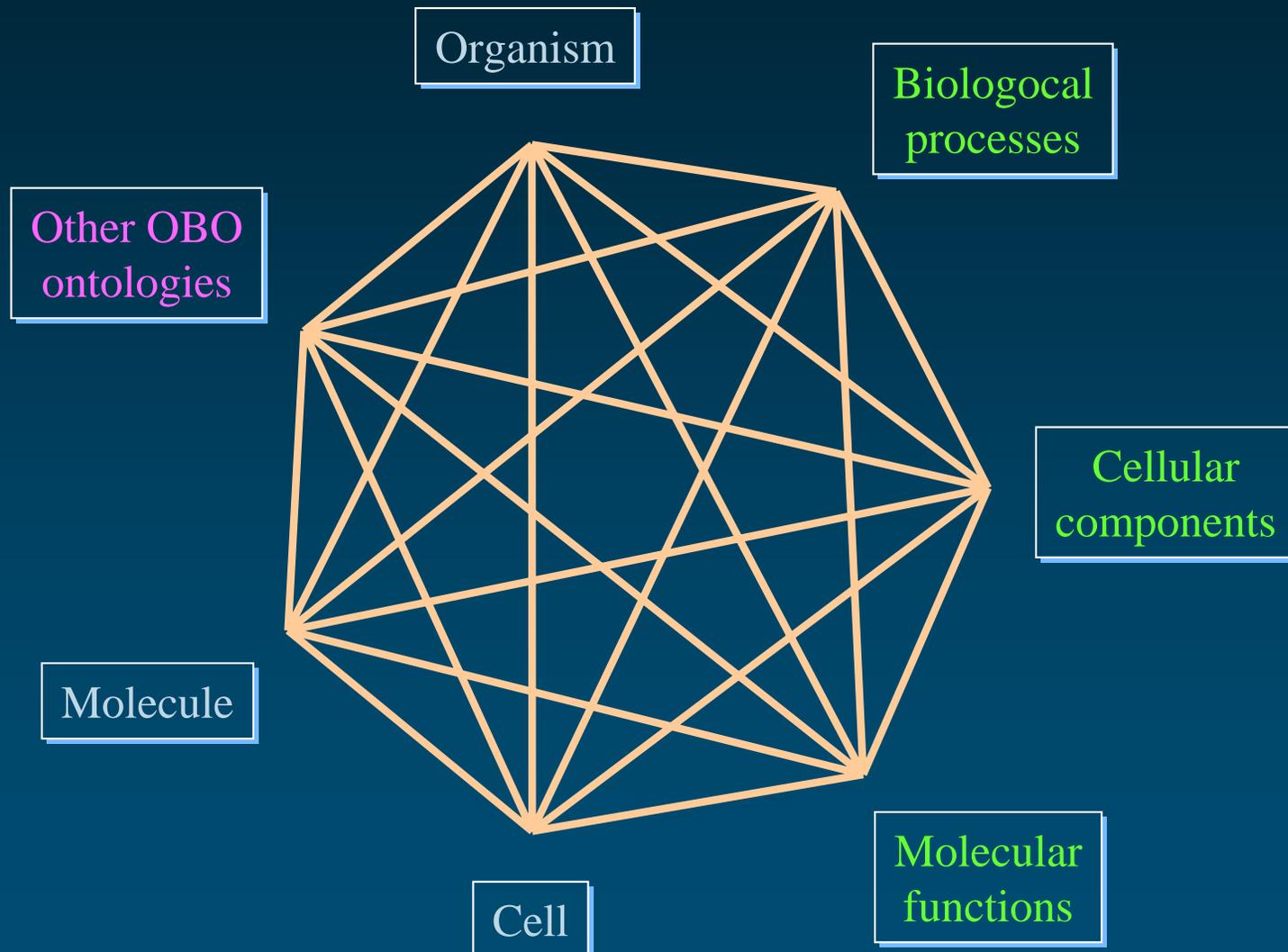
GO and other domains (revisited)

Resolution	Physical whole	Physical part	Function	Process
	Organism	Organism components	Organism functions	Organism processes
	Cell	Cellular components	Cellular functions	Cellular processes
	Molecule	Molecular components	Molecular functions	Molecular processes

[adapted from B. Smith]



Integrating biological ontologies



Linking GO to ChEBI

[Burgun & al., SMBM 2005]

ChEBI

- ◆ Member of the OBO family
- ◆ Ontology of
Chemical **E**ntities of **B**iological **I**nterest
 - Atom
 - Molecule
 - Ion
 - Radical
- ◆ 10,516 entities
 - 27,097 terms [Dec. 22, 2004]



Methods

- ◆ Every ChEBI term searched in every GO term
- ◆ Maximize precision
 - Ignored ChEBI terms of 3 characters or less
 - Proper substring
- ◆ Maximize recall
 - Case insensitive matches
 - Normalized ChEBI names
(generated singular forms from plurals)



Examples

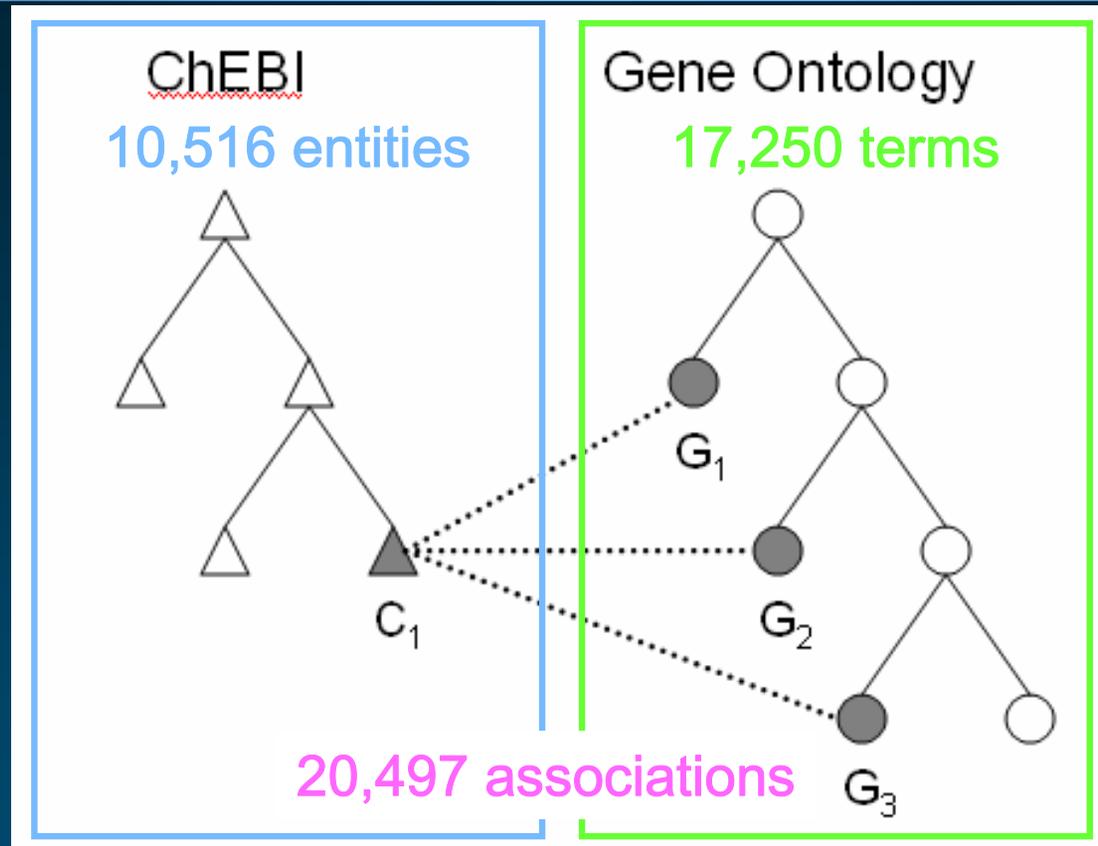
- ◆ **iron** [CHEBI:18248]
 - BP iron ion transport [GO:0006826]
 - MF iron superoxide dismutase activity [GO:0008382]
 - CC vanadium-iron nitrogenase complex [GO:0016613]

- ◆ **uronic acid** [CHEBI:27252]
 - BP uronic acid metabolism [GO:0006063]
 - MF uronic acid transporter activity [GO:0015133]

- ◆ **carbon** [CHEBI:27594]
 - BP response to carbon dioxide [GO:0010037]
 - MF carbon-carbon lyase activity [GO:0016830]



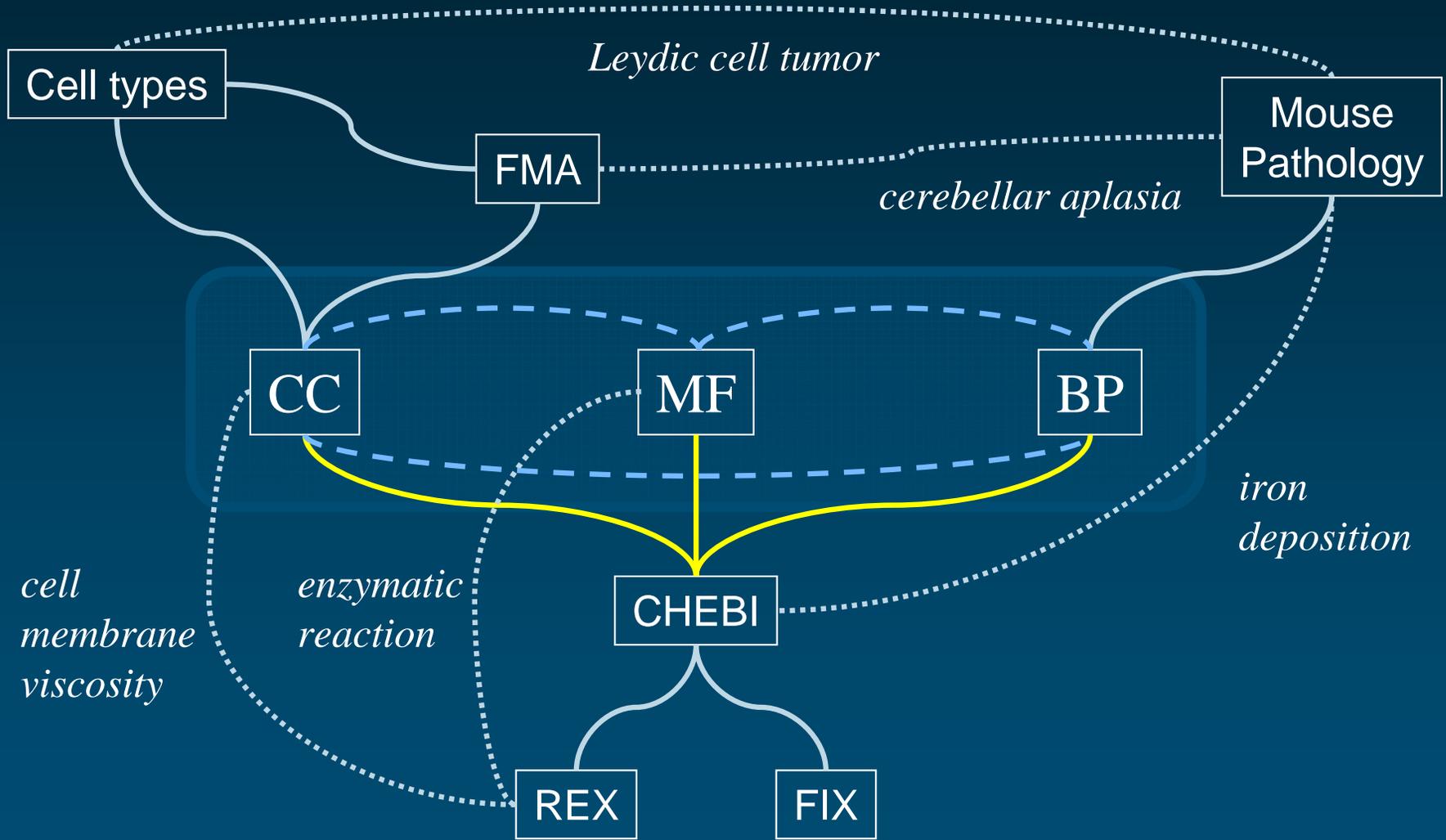
Quantitative results



- ◆ 2,700 ChEBI entities (27%) identified in some GO term

- ◆ 9,431 GO terms (55%) include some ChEBI entity in their names

Generalization



Conclusions

Conclusions (1)

- ◆ Links across OBO ontologies need to be made explicit
 - Between GO terms across GO hierarchies
 - Between GO terms and OBO terms
 - Between terms across OBO ontologies
- ◆ Automatic approaches
 - Effective (GO-GO, GO-ChEBI)
 - At least to bootstrap the process
 - Needs to be refined

Conclusions (2)

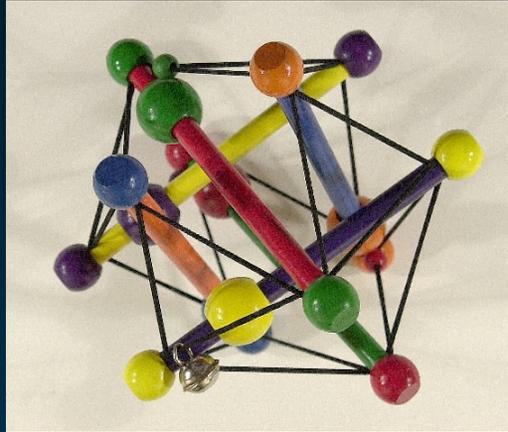
- ◆ Affordable relations
 - Computer-intensive, not labor-intensive
- ◆ Methods must be combined
 - Cross-validation
 - Redundancy as a surrogate for reliability
 - Relations identified specifically by one approach
 - False positives
 - Specific strength of a particular method
- ◆ Requires (some) manual curation
 - Biologists must be involved



References

- ◆ Bodenreider O, Aubry M, Burgun A. *Non-lexical approaches to identifying associative relations in the Gene Ontology*. In: Altman RB, Dunker AK, Hunter L, Jung TA, Klein TE, editors. Pacific Symposium on Biocomputing 2005: World Scientific; 2005. p. 91-102.
<http://mor.nlm.nih.gov/pubs/pdf/2005-psb-ob.pdf>
- ◆ Burgun A, Bodenreider O. *An ontology of chemical entities helps identify dependence relations among Gene Ontology terms*. Proceedings of the First International Symposium on Semantic Mining in Biomedicine (SMBM-2005)
Electronic proceedings: CEUR-WS/Vol-148
<http://mor.nlm.nih.gov/pubs/pdf/2005-smbm-ab.pdf>





Medical Ontology Research

Contact: olivier@nlm.nih.gov

Web: mor.nlm.nih.gov



Olivier Bodenreider

Lister Hill National Center
for Biomedical Communications
Bethesda, Maryland - USA