Enhancing Ontologies Through Annotations

Olivier Bodenreider

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

Outline

- Dependence relations in MeSH and co-occurrence in MEDLINE
- Identifying associative relations in the Gene Ontology
- Linking the Gene Ontology to other biological ontologies: GO-ChEBI

Using Dependence Relations in MeSH as a Framework for the Analysis of Disease Information in Medline

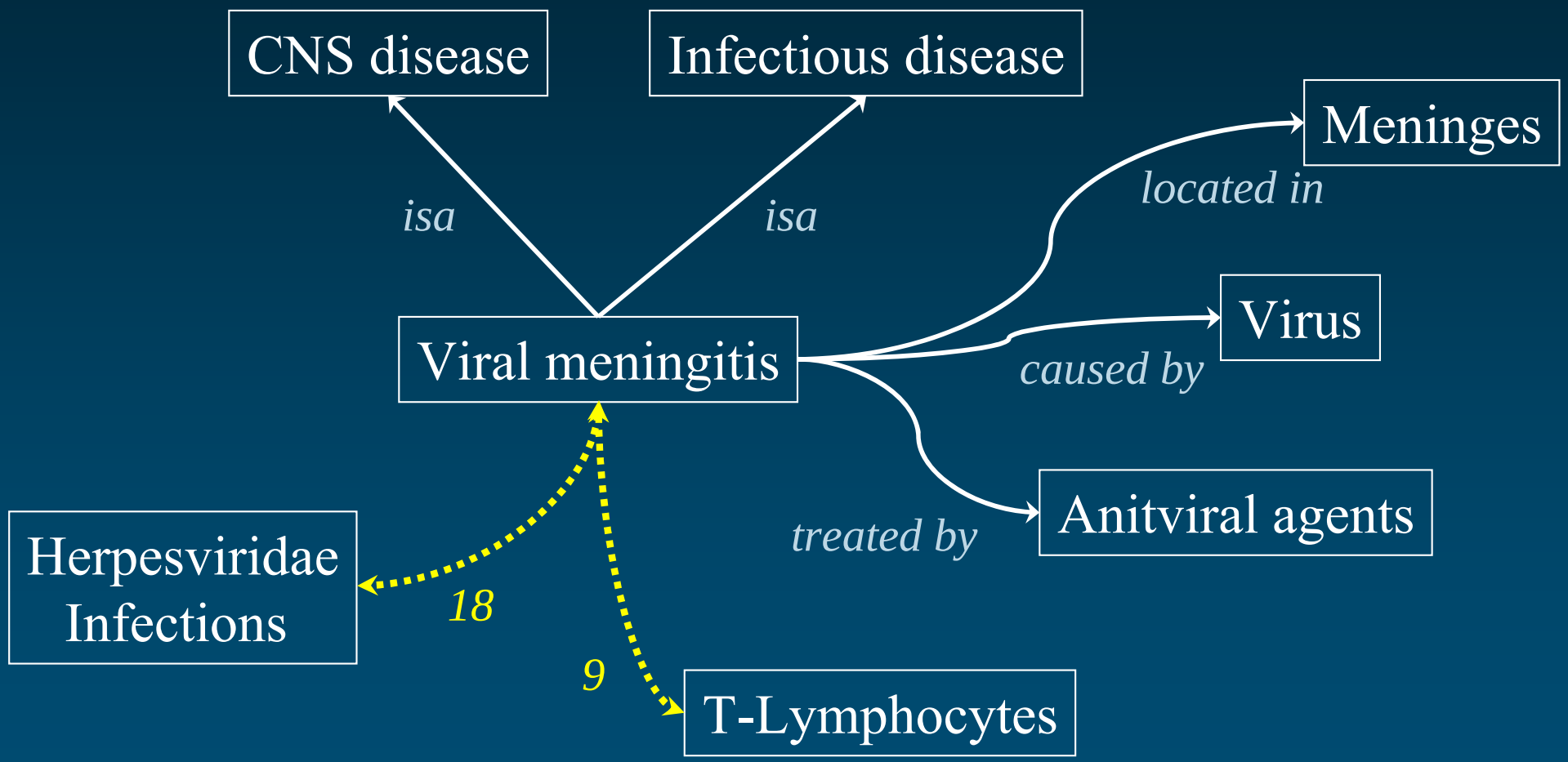
Acknowledgments

- Lowell Vizenor
  National Library of Medicine, USA

Relations among biomedical entities

- Symbolic relations
  - Represented in biomedical terminologies/ontologies
  - Explicit semantics
    - Hierarchical (*isa*, *part of*)
    - Associative (*location of*, *causes*, …)
- Statistical relations
  - Represented in text
    - Among lexical items (entity recognition)
    - Annotations
  - No explicit semantics

Example Viral meningitis



Statistical relations

- Crucial for text mining applications
  - Entity recognition
  - Frequency of co-occurrence

- No semantics
- Frequency of co-occurrence used as an indicator of the salience of the relation

An example from MEDLINE

Hurwitz JL, Korngold R, Doherty PC.
Specific and nonspecific T-cell recruitment in viral meningitis: possible implications for autoimmunity.
Cell Immunol. 1983 Mar;76(2):397-401.

Specific and nonspecific T-cell invasion into cerebrospinal fluid has been investigated in the nonfatal viral meningoencephalitis induced by intracerebral inoculation of mice with vaccinia virus. At the peak of the inflammatory process on Day 7 approximately 5 to 10% of the Lyt 2+ T cells present are apparently specific for vaccinia virus. Concurrently, in mice primed previously with influenza virus, 0.5 to 1.0% of the appropriate T-cell set located in cerebrospinal fluid is reactive to influenza-infected target cells. This vaccinia virus-induced inflammatory exudate may thus contain as many as 500 influenza-immune memory T cells. These findings are discussed from the aspect that such nonspecific T-cell invasion into the central nervous system during aseptic viral meningitis could result in exposure of potentially brain-reactive T cells to central nervous system components. PMID: 6601524

- Brain/immunology
- Cytotoxicity, Immunologic
- Exudates and Transudates/cytology
- Exudates and Transudates /immunology
- Meningitis, Viral/immunology*
- T-Lymphocytes/immunology*
- Vaccinia virus

- Animals
- Humans
- Mice
- Research Support, Non-U.S. Gov't
- Research Support, U.S. Gov't, P.H.S.

Ontological analysis

- Formal ontological distinction
  - Dependence relations
    - Every instance of a class is related to some instances of another class
    - *A* is ontologically dependent on *B* if and only if *A* exists then *B* exists
  - Contingent relations
    - Only some instances of a class are related to some instances of another class

CNS disease → CNS

Viral meningitis → Virus

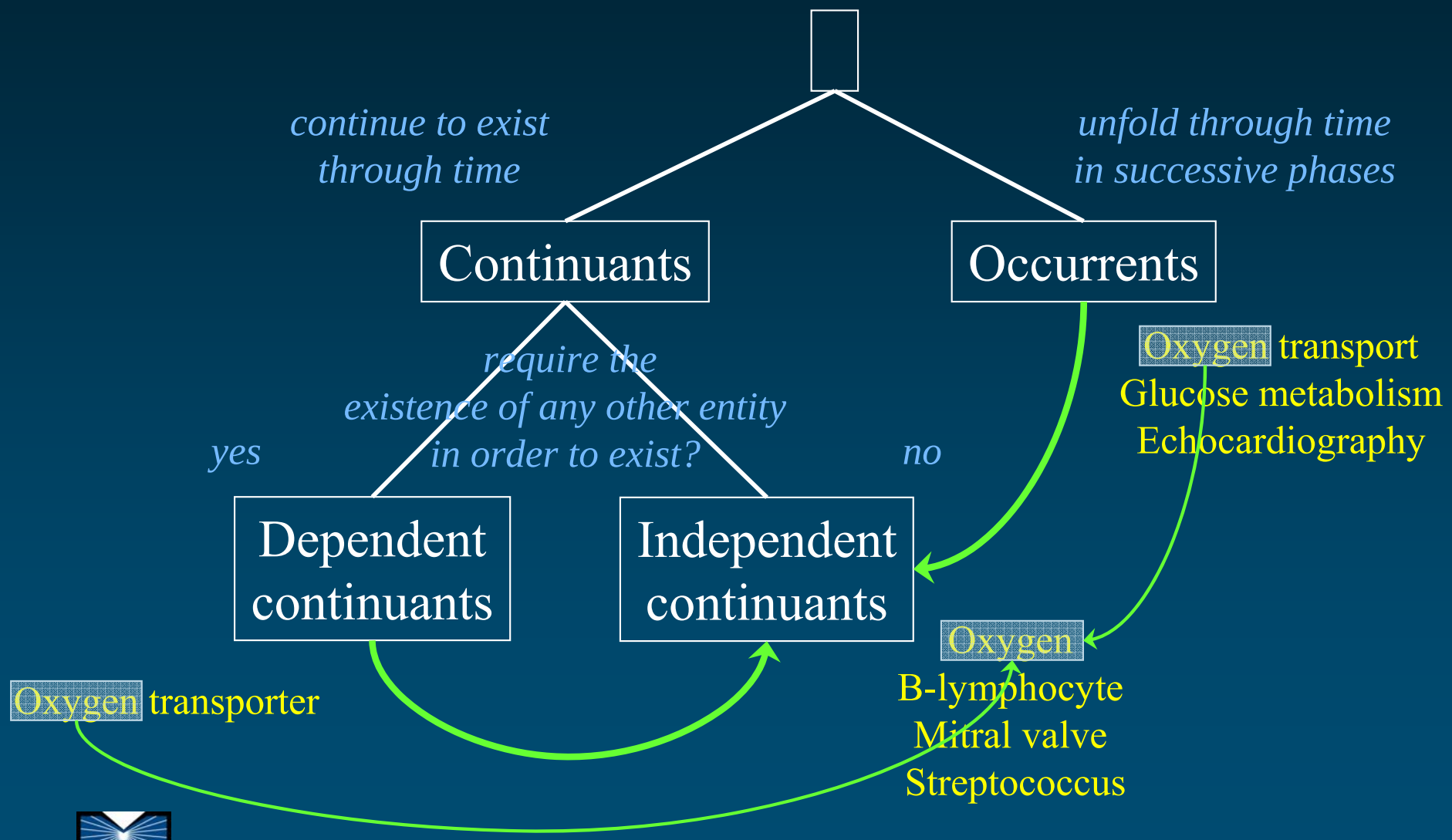
aspirin - - - → headache

Statistical vs. ontological

- Can we use formal ontology to help analyze statistical relations?
- What is the relation between ontological and statistical relations?

- Hypothesis:
  - Correspondence between
    - Dependence relations (ontological)
    - High frequencies of co-occurrence (statistical)
  - Dependence relations ↔ Systematically high frequencies of co-occurrence

More formal-ontological distinctions



Application to diseases

- Diseases are (mostly) processes, i.e., occurrents
- Diseases are dependent entities
- Diseases depend on independent continuants
  - Anatomical structures
    - Classification by “location” (body system)
  - Agents (pathogens)
    - Classification by etiology

Participation relation

- Participation relations are dependence relations
- Between processes and biomedical continuants
- Passive participation: *has_participant*
  - Viral meningitis *has_participant* Meninges
- Active participation: *has_agent*
  - Viral meningitis *has_agent* Virus

- Defined at the instance level
  but can be adapted at the class level

Statistical relations

- Independent events
  - $P(E1 \cap E2) = P(E1) \,.\, P(E2)$
- Tests of independence
  - χ^2 test
  - G^2 test (likelihood ratio test)

Objectives

- Analyze dependence relations in MeSH and to compare them to statistical relations obtained from co-occurrence data
- Restricted to the relations between disease categories and other categories of biomedical interest

- Hypothesis:
  - Co-occurrence relations between diseases and other categories
    - Highest proportion for the dependent category, systematically across diseases
    - Smaller proportions for other non-dependent categories

Materials

Medical Subject Headings (MeSH)

- Controlled vocabulary used to index MEDLINE
- 22,658 descriptors (2004 version)
- 16 tree-like hierarchies
  - Anatomy
  - Organisms
  - Diseases
  - …

1. Anatomy [A]
2. Organisms [B]
3. Diseases [C]
4. Chemicals and Drugs [D]
5. Analytical, Diagnostic and Therapeutic Techniques and Equipment [E]
6. Psychiatry and Psychology [F]
7. Biological Sciences [G]
8. Physical Sciences [H]
9. Anthropology, Education, Sociology and Social Phenomena [I]
10. Technology and Food and Beverages [J]
11. Humanities [K]
12. Information Science [L]
13. Persons [M]
14. Health Care [N]
15. Publication Characteristics [V]
16. Geographic Locations [Z]

MEDLINE

- 385,491 citations (year 2004)
- Indexed with 20,085 distinct MeSH descriptors

- Restrictions
  - Starred descriptors only (3.5 / citation, on average)
  - Frequency of co-occurrence $\geq$ 10
  - Associations between diseases and other categories

Methods and Results

3. ⊟ Diseases [C]

- Bacterial Infections and Mycoses [C01] +
- Virus Diseases [C02] +
- Parasitic Diseases [C03] +
- Neoplasms [C04] +
- Musculoskeletal Diseases [C05] +
- Digestive System Diseases [C06] +
- Stomatognathic Diseases [C07] +
- Respiratory Tract Diseases [C08] +
- Otorhinolaryngologic Diseases [C09] +
- Nervous System Diseases [C10] +
- Eye Diseases [C11] +
- Urologic and Male Genital Diseases [C12] +
- Female Genital Diseases and Pregnancy Complications [C13] +
- Cardiovascular Diseases [C14] +
- Hemic and Lymphatic Diseases [C15] +
- Congenital, Hereditary, and Neonatal Diseases and Abnormalities [C16] +
- Skin and Connective Tissue Diseases [C17] +
- Nutritional and Metabolic Diseases [C18] +
- Endocrine System Diseases [C19] +
- Immune System Diseases [C20] +
- Disorders of Environmental Origin [C21] +
- Animal Diseases [C22] +
- Pathological Conditions, Signs and Symptoms [C23] +

Identifying dependence relations

- Manual examination of the 23 top-level disease categories [C tree]
  - Exceptions
    - *Pathological conditions, signs and symptoms* (C23)
  - Additions
    - *Mental disorders* (F03)
- Identify the categories in (active or passive) participation relation with the process

Identifying dependence relations Results

has_participant

Pathological process	Anatomical entity
Musculoskeletal Diseases	Musculoskeletal System
Digestive System Diseases	Digestive System
Stomatognathic Diseases	Stomatognathic System
Respiratory Tract Diseases	Respiratory System
Nervous System Diseases	Nervous System
Eye Diseases	Sense Organs (+)
Urological and Male Genital Diseases	Urogenital System
Female Genital Diseases and Pregnancy Complications	Urogenital System Embryonic Structures
Cardiovascular Diseases	Cardiovascular System
Hemic and Lymphatic Diseases	Hemic and Immune Systems
Skin Diseases	Integumentary System
Endocrine Diseases	Endocrine System

Identifying dependence relations Results

has_agent

Pathological Process	Pathogen
Bacterial Infection and Mycoses	Bacteria Fungi
Virus Diseases	Viruses
Parasitic Diseases	Animals (+)

Identifying statistical relations

- Aggregation at the level of top-level categories

Nervous System Diseases [C10]
Central Nervous System Diseases [C10.228]
Central Nervous System Infections [C10.228.228]
Central Nervous System Viral Diseases [C10.228.228.210]
Encephalitis [C10.228.228.210.150] +
Meningitis, Viral [C10.228.228.210.500]

Virus Diseases [C02]
Central Nervous System Viral Diseases [C02.182]
Encephalitis [C02.182.500] +
Meningitis, Viral [C02.182.550]

Nervous System [A08]
Central Nervous System [A08.186]
Brain [A08.186.211] +
Meninges [A08.186.566]

Viral meningitis (C02, C10)

(C02, A08)
and
(C10, A08)

Meninges (A08)

Identifying statistical relations

- Contingency table

Indexed with term *B*

Indexed with term *A*

	Yes	No
Yes	n_{AB}	n_{Ab}
No	n_{aB}	n_{ab}

- Testing independence
  - G^2 test (likelihood ratio test)

Identifying statistical relations Results

- Quantitative results
  - 25,376 pairs of co-occurring descriptors
  - All but 68 of these statistically significant (G^2 test)
  - 7,896 pairs with frequency of co-occurrence ≥ 10
  - 6,525 between diseases and other categories

Identifying statistical relations Results

23 disease top-level categories

88 other top-level categories

		C01 Bacterial Infections and Mycoses	C02 Virus Diseases	C03 Parasitic Diseases	C04 Neoplasms	C05 Musculoskeletal Diseases	C06 Digestive System Diseases	C07 Stomatognathic Diseases	C08 Respiratory Tract Diseases
A01	Body Regions	0.00	0.00	0.00	6.75	1.00	3.50	0.00	0.00
A02	Musculoskeletal System	1.50	0.00	0.00	6.17	52.08	0.00	2.58	1.75
A03	Digestive System	1.50	1.50	0.00	18.83	0.00	45.75	0.33	0.17
A04	Respiratory System	0.50	0.50	0.00	3.58	0.25	1.33	0.00	21.25
A05	Urogenital System	0.00	1.00	0.00	9.08	0.00	1.00	0.00	0.00
A06	Endocrine System	0.00	0.00	0.00	3.50	0.00	0.00	0.00	0.00
A07	Cardiovascular System	1.00	0.00	0.00	5.17	0.00	2.50	0.00	0.50

Identifying statistical relations Results

- Qualitative results (1)
  - Generally one top-level category of the *Anatomy* and *Organisms* trees accounting for the highest frequency of co-occurrence for a given disease
    - *Cardiovascular Diseases* → *Cardiovascular System*
  - Exceptions
    - *Neoplasms* [C04]
    - *Congenital, Hereditary, and Neonatal Diseases and Abnormalities* [C16]
    - *Endocrine Diseases* [C19]
    - *Immunologic Diseases* [C20]

Identifying statistical relations Results

- Qualitative results (2)
  - Most *Anatomy* and *Organisms* categories are preferentially associated with one disease category
    - *Cardiovascular System* → *Cardiovascular Diseases*
  - Categories other than *Anatomy* and *Organisms* tend not to be associated with one particular disease category (contingent rather than dependent relations)
    - *Pathological Conditions, Signs and Symptoms* [C23]
    - *Amino Acids, Peptides, and Proteins* [D12]
    - *Diagnosis* [E01]
    - *Therapeutics* [E02]
    - *Surgical Procedures, Operative* [E04]

Discussion

Applications

- To semantic mining
  - Formal ontological analysis of relations provides a useful framework for elucidating statistical associations
- To terminology creation and maintenance
  - Most terminologies do not represent trans-ontological relations explicitly
  - Concepts in dependence relation should not be modified independently of each other

Summary

- We have studied statistical associations between MeSH terms co-occurring in MEDLINE citations
- We have shown that the ontological relation of dependence is generally corroborated by a strong, systematic statistical association
- These techniques
  - Provide a framework for semantic mining of diseases
  - Can help maintain terminologies

References

- Vizenor L, Bodenreider O. *Using dependence relations in MeSH as a framework for the analysis of disease information in Medline.* Proceedings of the Second International Symposium on Semantic Mining in Biomedicine (SMBM-2006) 2006:76-83. http://mor.nlm.nih.gov/pubs/pdf/2006-smbm-lv.pdf

Non-lexical Approaches to Identifying Associative Relations in the Gene Ontology

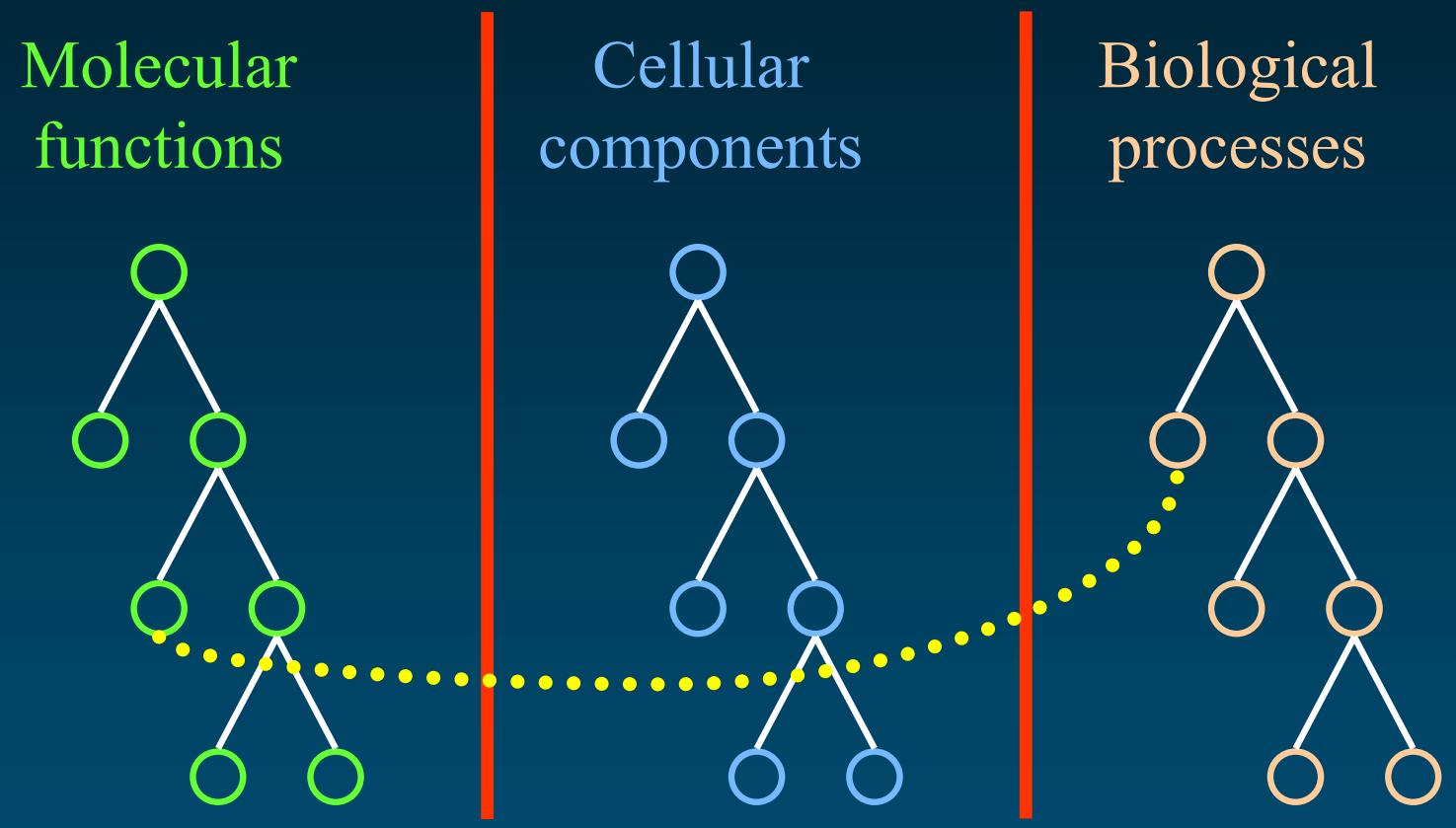
Acknowledgments

- 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



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]

GO and annotation databases

- 5 model organisms
  - FlyBase
  - GOA-Human
  - MGI
  - SGD
  - WormBase

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	…	

270,000 gene-term associations

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

❶ Similarity in the vector space model

GO terms

Genes

	t_1	t_2			…				t_n
g_1	●				●			●	
g_2		●					●		●
	●				●			●	
…				●		●			
g_n			●		●		●		●

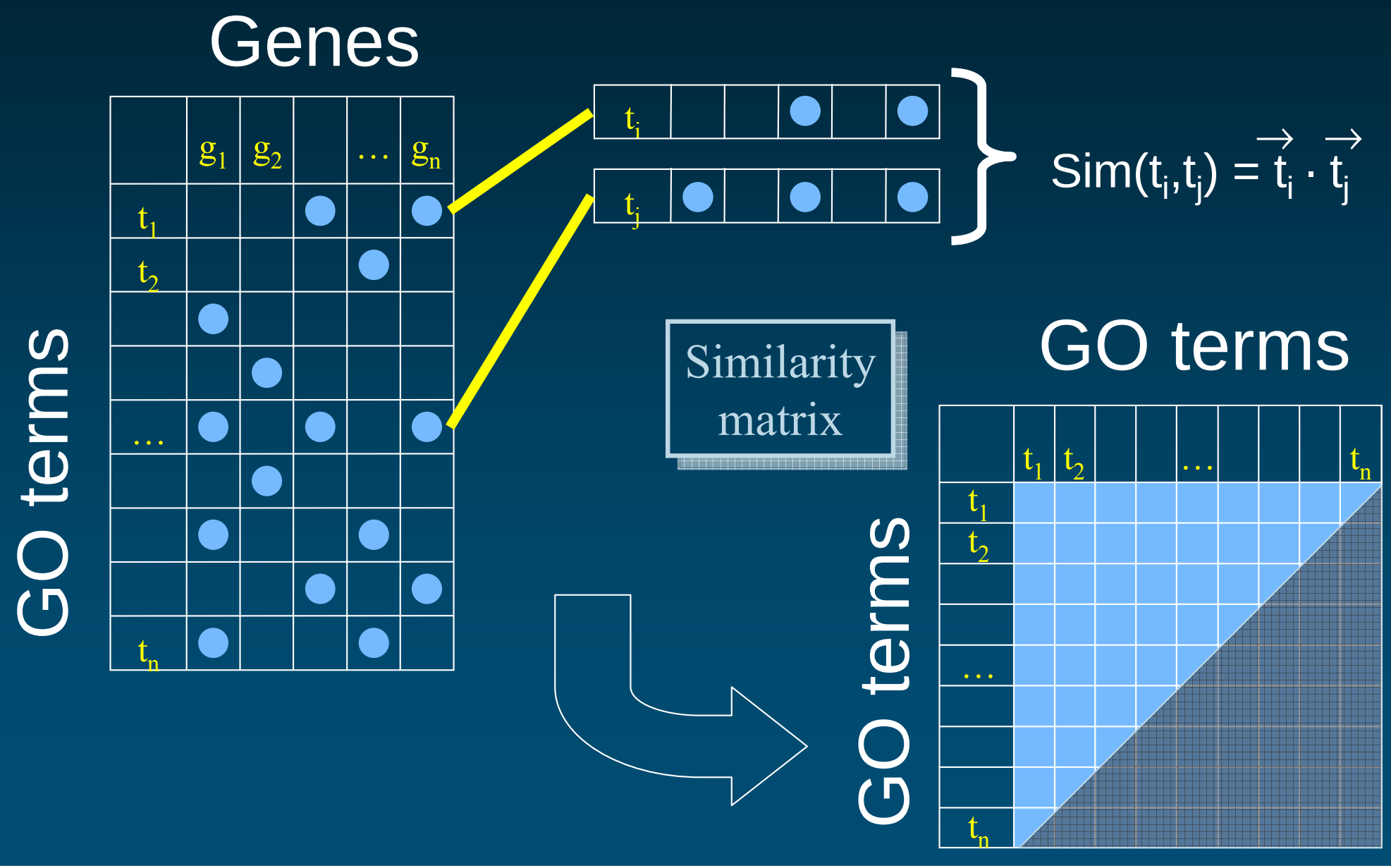
Annotation database

Genes

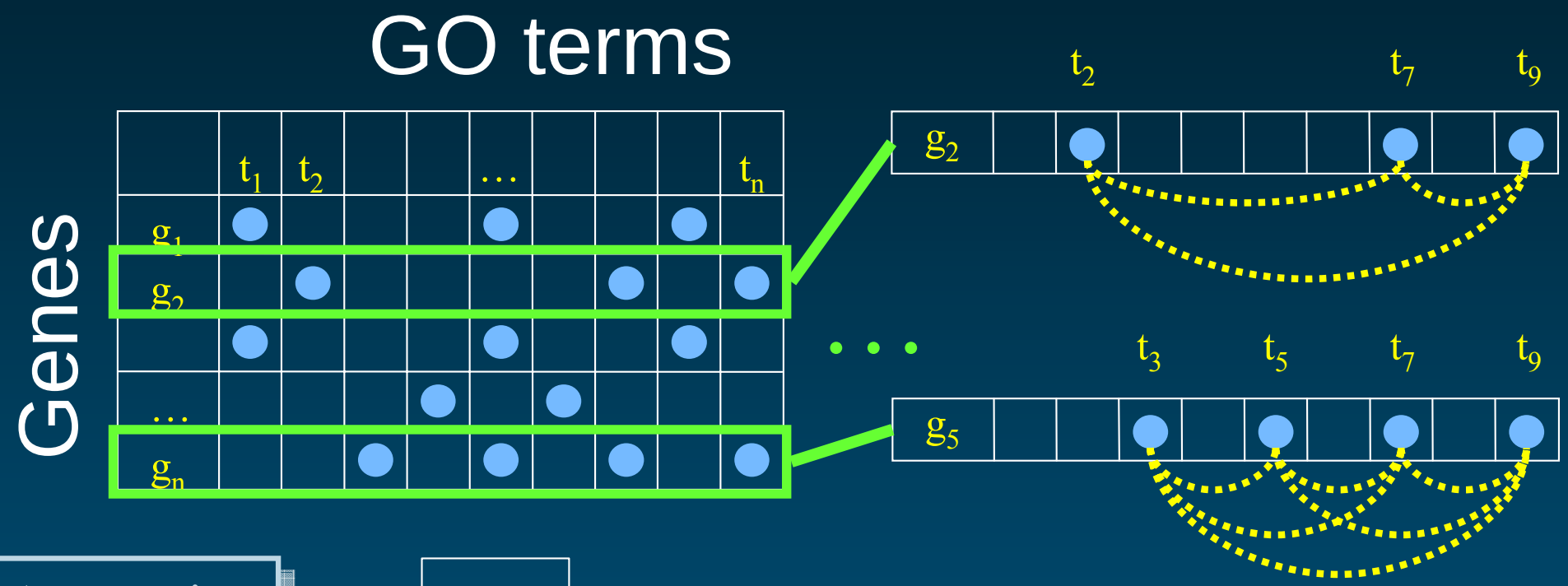
GO terms

	g_1	g_2		…	g_n
t_1			●		●
t_2				●	
	●				
		●			
…	●		●		●
		●			
	●			●	
			●		●
t_n	●			●	

❶ Similarity in the vector space model



❷ 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
…	

❷ 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*

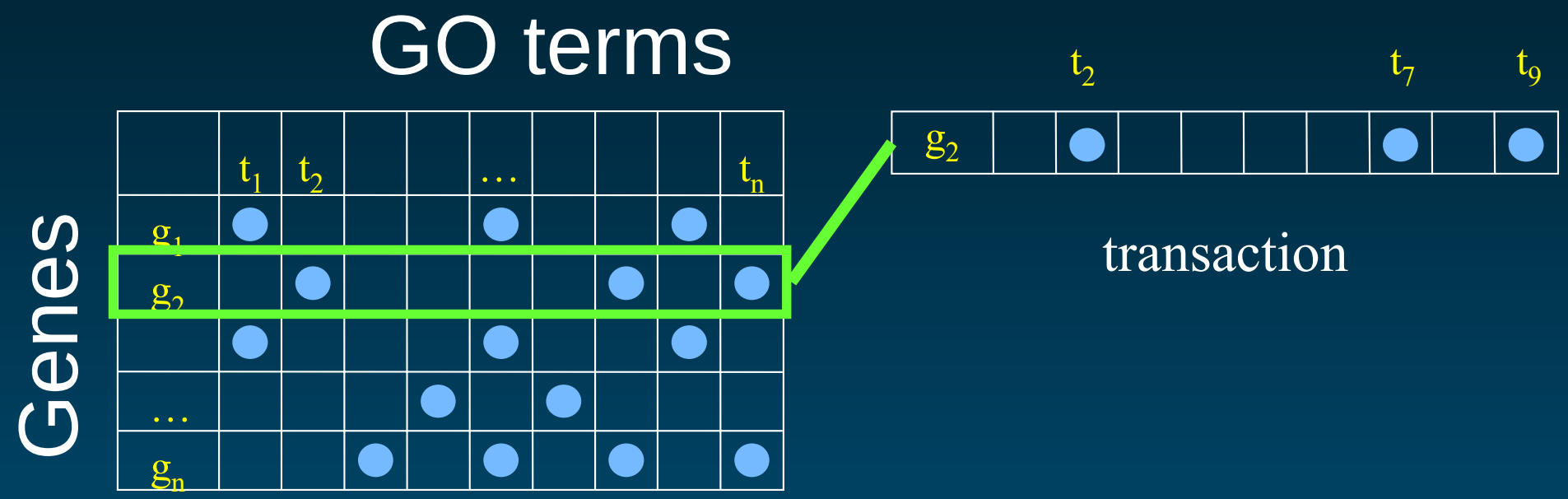
GO:0006955 *chemokine activity*

	present	absent	Total
present	46	542	588
absent	7	21,583	22,132
total	53	22,125	22,720

$G^2 = 298.7$

$p < 0.000$

❸ Association rule mining



Annotation database

apriori

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

Examples of associations

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

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

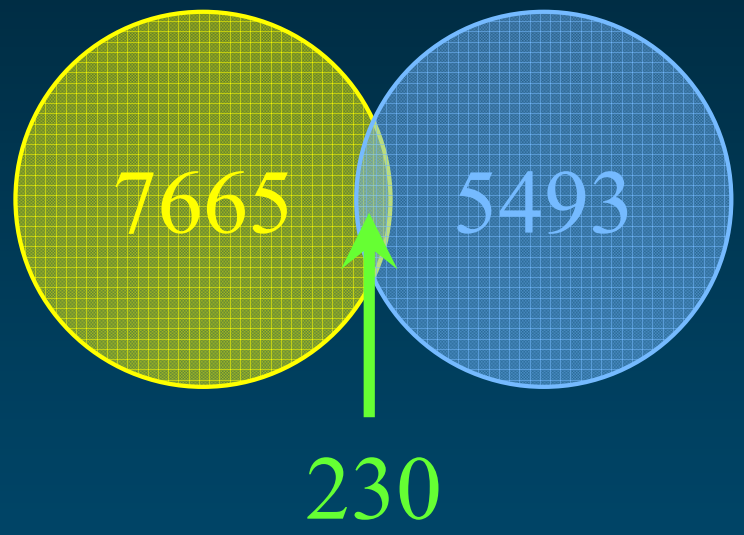
Associations identified ARM

	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

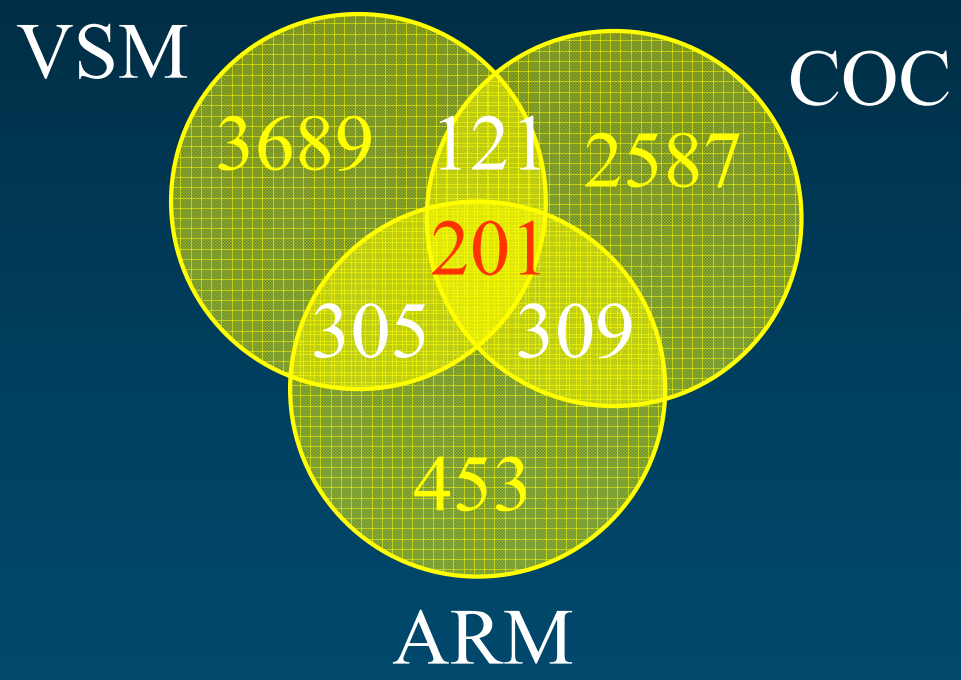
MF: carboxypeptidase A activity
BP: peptolysis and peptidolysis

Limited overlap among approaches

- Lexical vs. non-lexical



- Among non-lexical



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

Linking the Gene Ontology to other biological ontologies

Acknowledgments

- Anita Burgun
  University of Rennes, France

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)

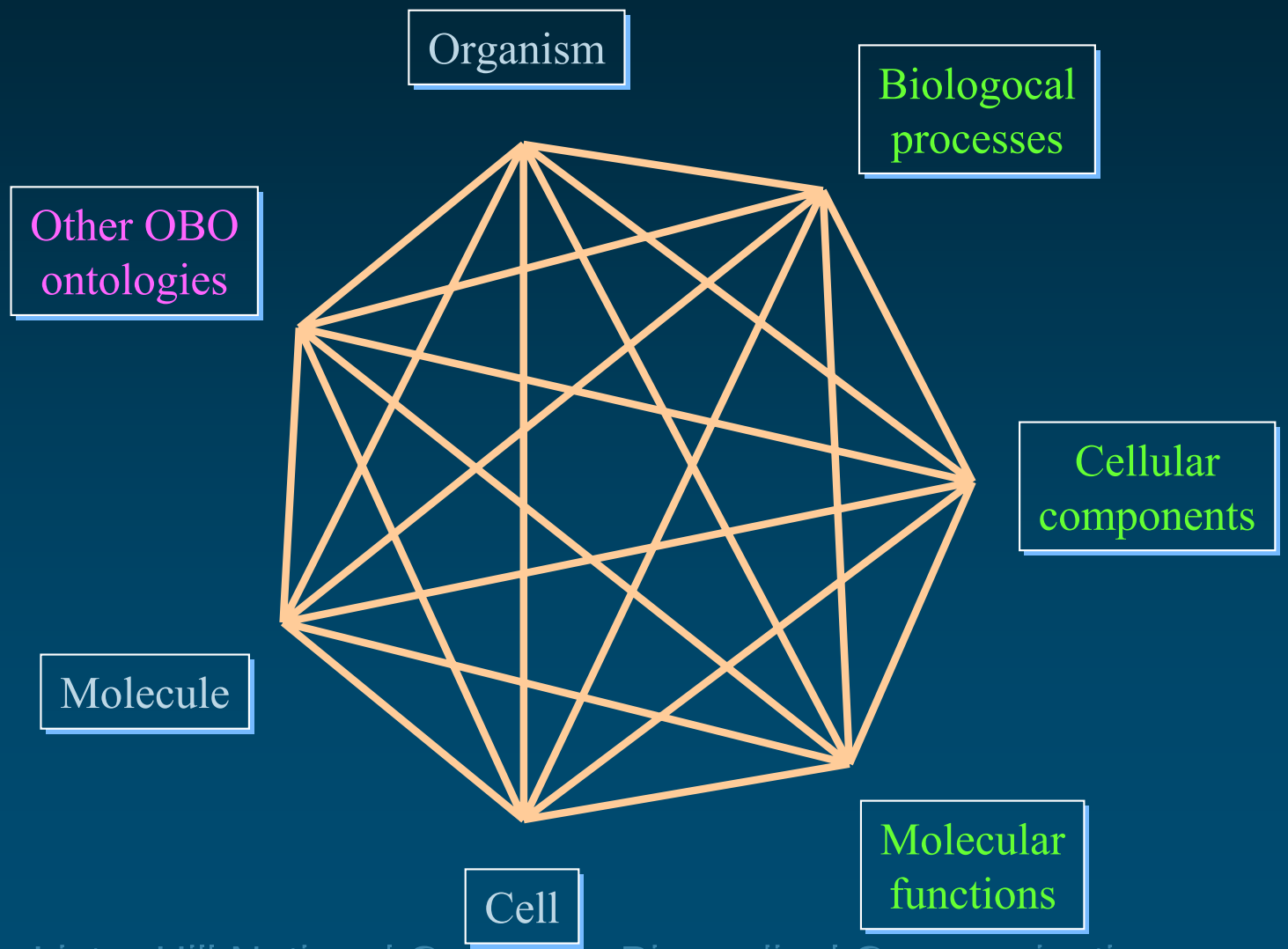
Domain	Prefix	Files
Cell type	CL	cell.obo
Chemical entities of biological interest	CHEBI	ontology.obo
Mus adult gross anatomy	MA	MA.ontology
Plant anatomy	PO	anatomy.ontology and anatomy.definition
NCBI organismal classification	taxon	taxonomy.dat
Human disease	DOID	DO 08 18 03.txt
Mouse pathology	MPATH	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

ChEBI

- Member of the OBO family
- Ontology of
  Chemical Entities of Biological Interest
  - Atom
  - Molecule
  - Ion
  - Radical
- 10,516 entities
  - 27,097 terms

[Dec. 22, 2004]

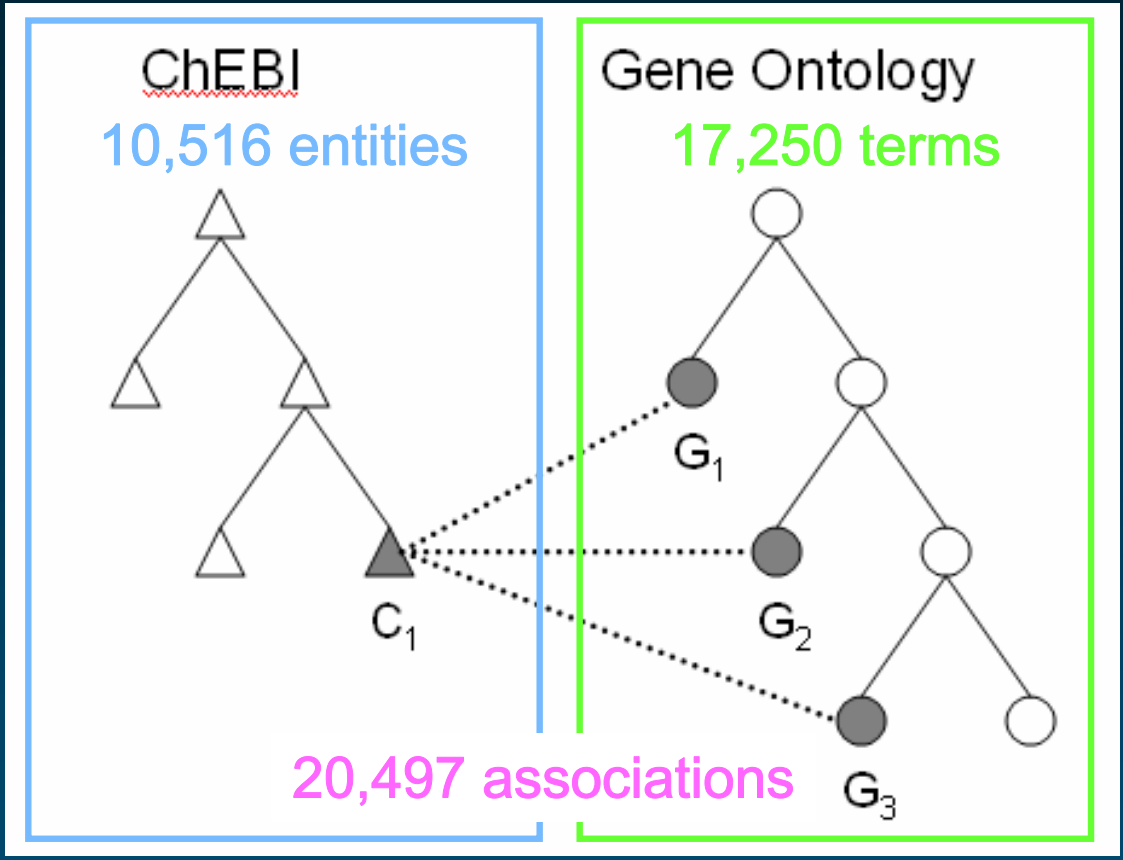
Methods

- 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

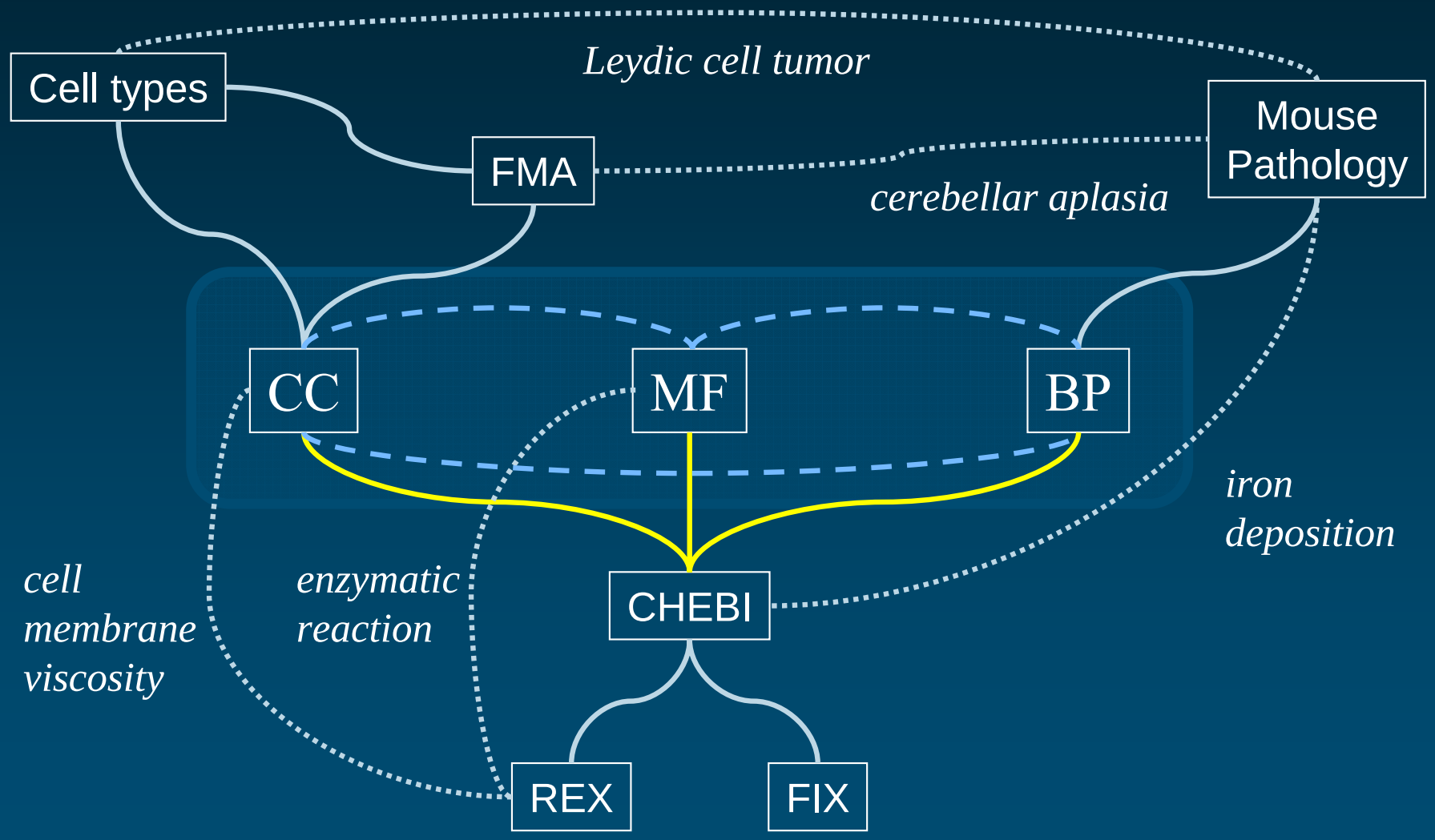
- iron [CHEBI:18248]
  - **BP** iron ion transport [GO:0006826]
  - **MF** iron superoxide dismutase activity [GO:0008382]
  - **CC** vanadium-iron nitrogenase complex [GO:0016613]
- uronic acid [CHEBI:27252]
  - **BP** uronic acid metabolism [GO:0006063]
  - **MF** uronic acid transporter activity [GO:0015133]
- 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)

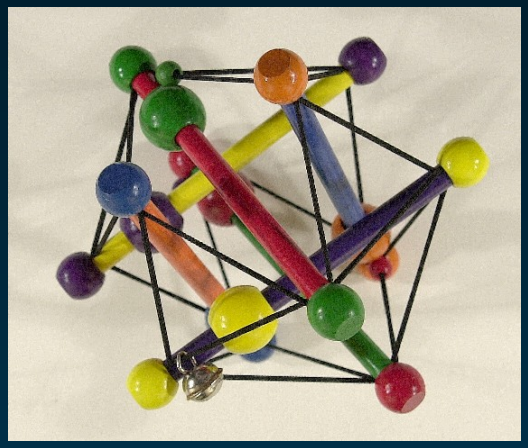
- Links across OBO ontologies need to be made explicit
  - Between GO terms across GO hierarchies
  - Between GO terms and OBO terms
  - Between terms across OBO ontologies
- Automatic approaches
  - Effective (GO-GO, GO-ChEBI)
  - At least to bootstrap the process
  - Needs to be refined

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

- 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