

Workshop on Ontological Foundations
of Biomedical Terminology Systems
October 22, 2005



Lexical and Statistical Approaches
to Acquiring Ontological Relations



Olivier Bodenreider

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

Introduction

◆ Biomedical ontologies

- Precisely defined (e.g., formal ontology)
- Limited size
- Built manually

◆ Large amounts of knowledge

- Not represented explicitly by symbolic relations
- But expressed implicitly
 - By lexico-syntactic relations (i.e., embedded in terms)
 - By statistical relations (e.g., co-occurrence)
- Can be extracted automatically



Ontology development

Formal ontology

- Provides a framework for building sound ontologies
- Too labor-intensive for building large ontologies

Otherwise

- Usually unsuitable for reasoning
- Tools for automatic acquisition available



General framework

- ◆ Ontology learning
 - [Maedche & Staab, Velardi]
 - ECAI, IJCAI
- ◆ Term variation [Jacquemin]
- ◆ Terminology / Knowledge TKE, TIA
- ◆ Knowledge acquisition/capture K-CAP
- ◆ Information extraction



Resources for ontology acquisition

- ◆ Long tradition of terminology building
 - Over 100 terminologies available in electronic format
- ◆ Large corpora available (e.g., MEDLINE)
 - Entity recognition tools available
 - E.g., MetaMap (UMLS-based)
 - Several for gene/protein names
 - Information extraction methods
- ◆ Large annotation databases available
 - MEDLINE citations indexed with MeSH
 - Model organism databases annotated with GO



Methods for ontology acquisition

◆ Lexico-syntactic methods

- Lexico-syntactic patterns
- Nominal modification
- Prepositional phrases
- Reified relations
- Semantic interpretation

◆ Statistical methods

- Clustering
- Statistical analysis of co-occurrence data
- Association rule mining



Lexico-syntactic methods

Compositional features of terms

- ◆ Lexical items [Baud & al., AMIA, 1998]
- ◆ Terms within a vocabulary
 - Clinical vocabularies [McDonald & al., AMIA, 1999]
 - Gene Ontology [Ogren & al., PSB, 2004]
[Mungall, CFG, 2004]
- ◆ Terms across vocabularies
 - SNOMED / LOINC [Dolin, JAMIA, 1998]
 - GO / ChEBI [Burgun, SMBM, 2005]
- ◆ Lexicon / Terms
 - Semantic lexicon [Johnson, JAMIA, 1999]
[Verspoor, CFG, 2005]



Statistical methods

Taxonomic relations Clustering

- ◆ Source: text corpus
- ◆ Principle: similarity between words reflected in their contexts
 - Co-occurring words (+ frequencies)
 - Hierarchical clustering algorithms
 - Similarity measure (cosine, Kullback Leibler)
- ◆ Can be refined using classification techniques (e.g., k nearest neighbors)

[Faure & al., LREC, 1998]

[Maedche & al., HoO, 2004]



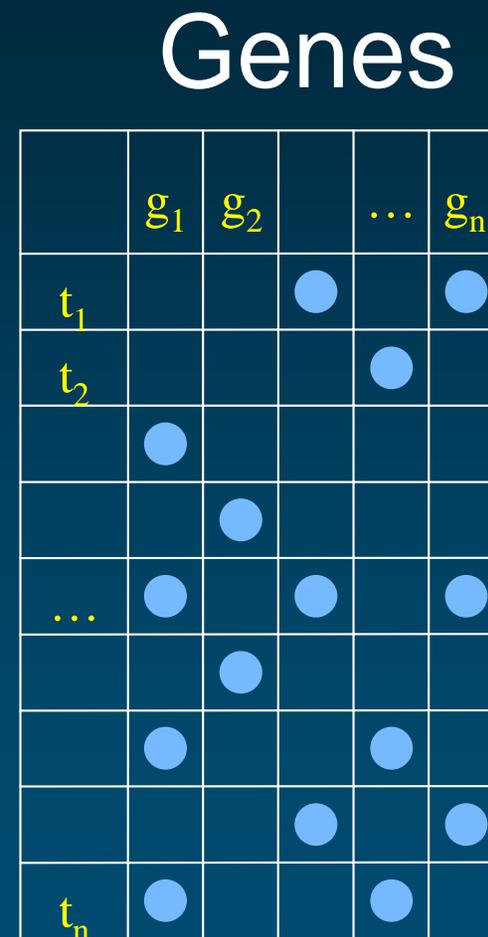
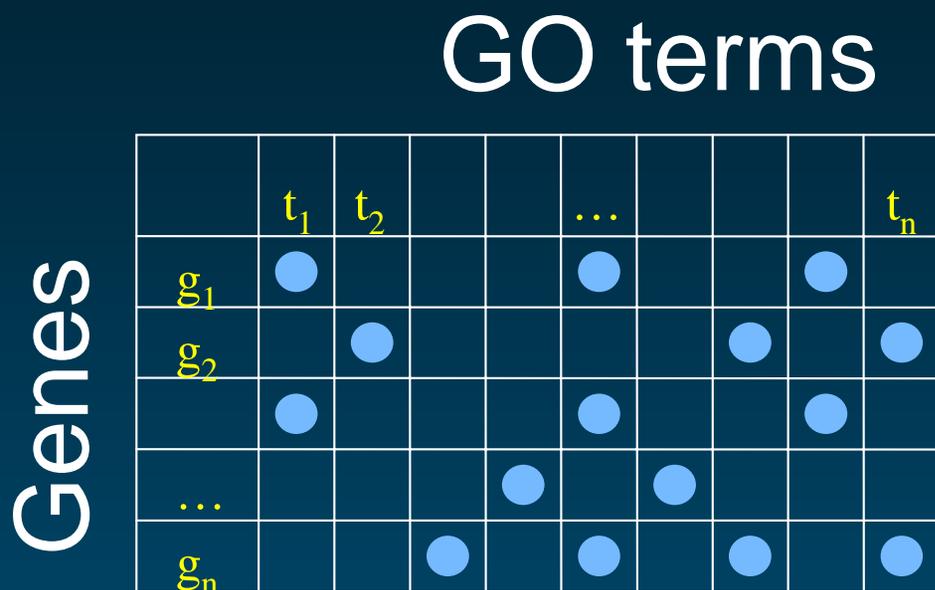
Associative relations

- ◆ Source: text corpus / annotation databases
- ◆ Principle: dependence relations
 - Associations between terms
- ◆ Several methods
 - Vector space model
 - Co-occurring terms
 - Association rule mining
- ◆ Limitations: no semantics

[Bodenreider & al., PSB, 2005]



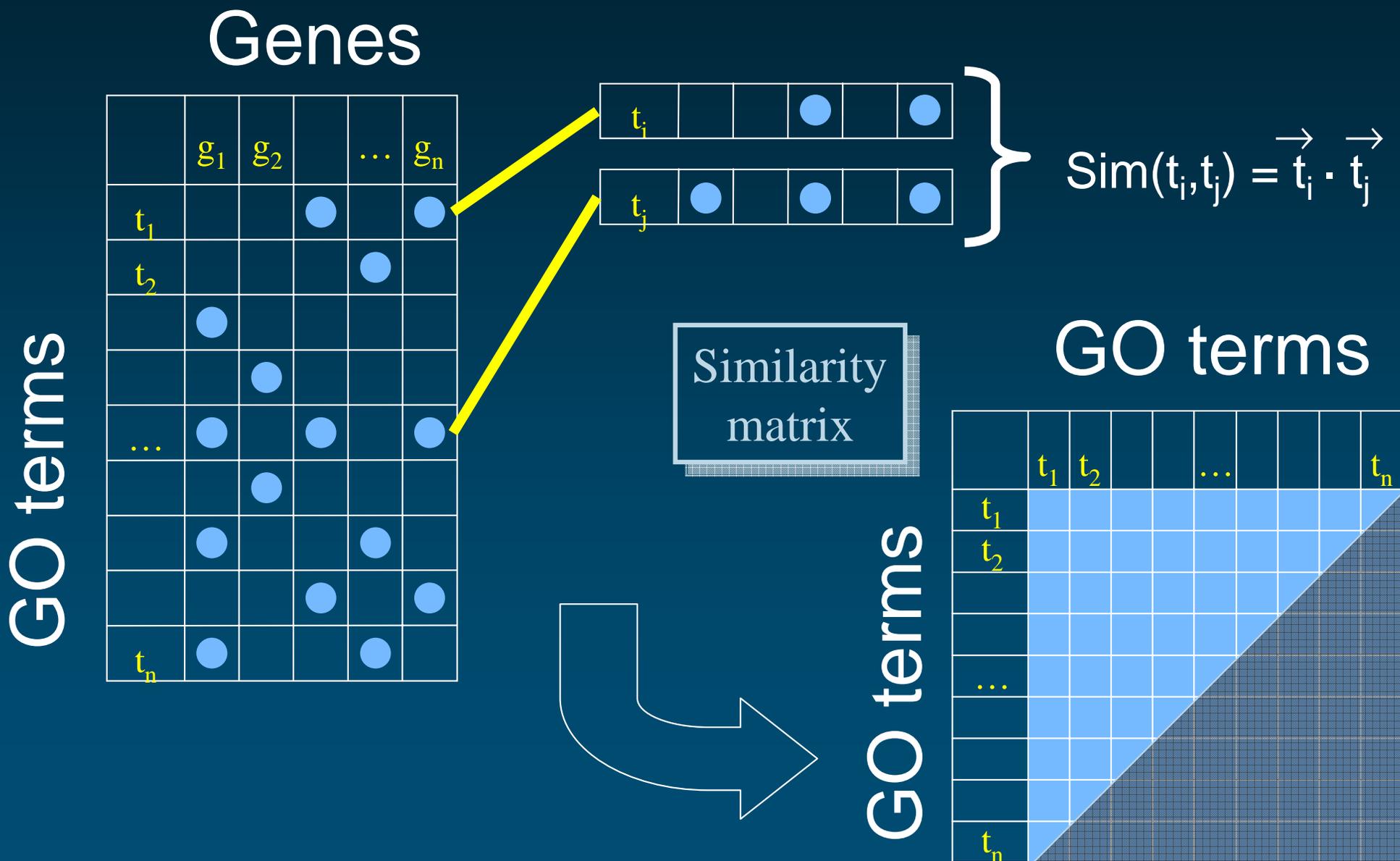
1 Similarity in the vector space model



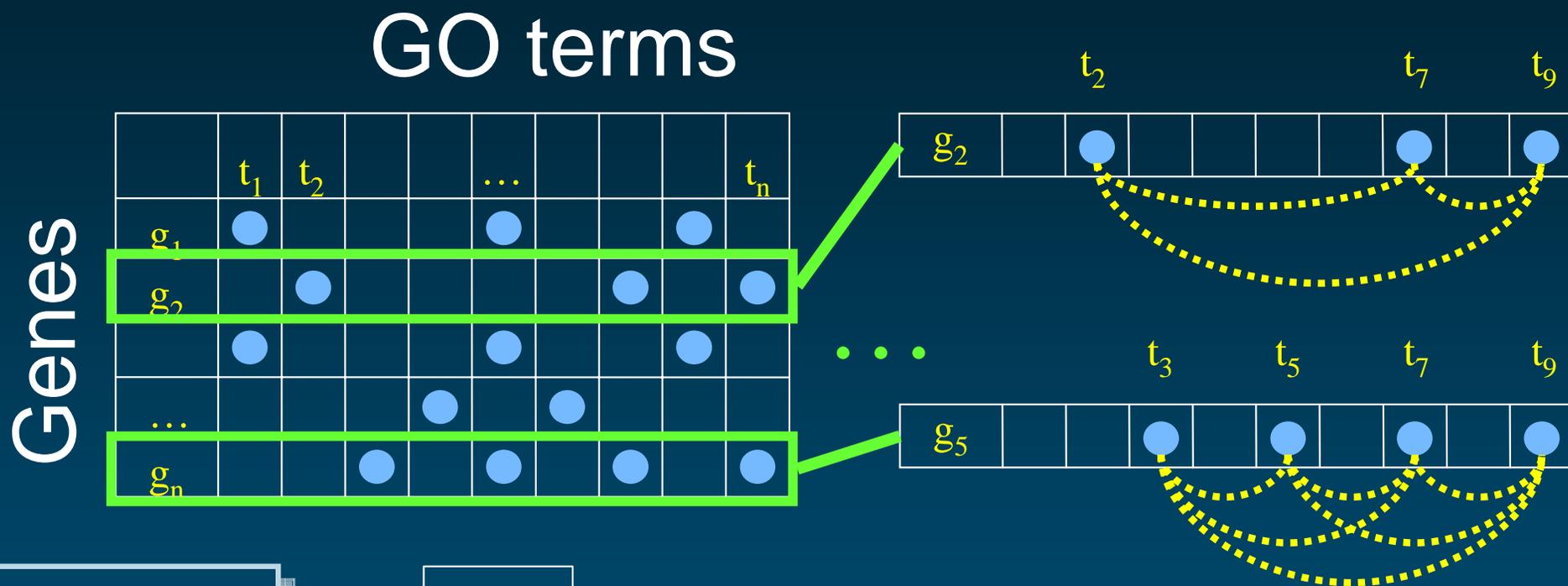
Annotation
database



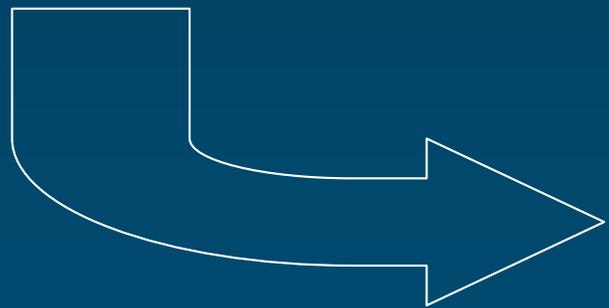
1 Similarity in the vector space model



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

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

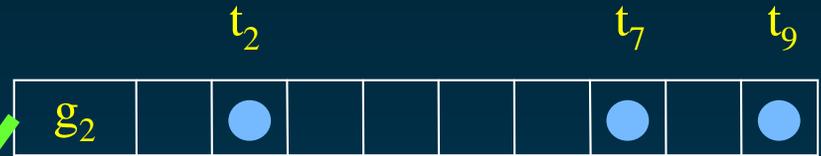
3

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: $.05$

Example of associations (GO)

- ◆ Vector space model
 - MF: *ice binding*
 - BP: *response to freezing*
- ◆ Co-occurring terms
 - MF: *chromatin binding*
 - CC: *nuclear chromatin*
- ◆ Association rule mining
 - MF: *carboxypeptidase A activity*
 - BP: *peptolysis and peptidolysis*



Discussion and Conclusions

Reusing thesauri

- ◆ First approximation for taxonomic relations
 - No need for creating taxonomies from scratch in biomedicine
- ◆ Beware of purpose-dependent relations
 - *Addison's disease* *isa* *Autoimmune disorder*
- ◆ Relations used to create hierarchies vs. hierarchical relations
- ◆ Requires (some) manual curation

[Wroe & al., PSB, 2003]

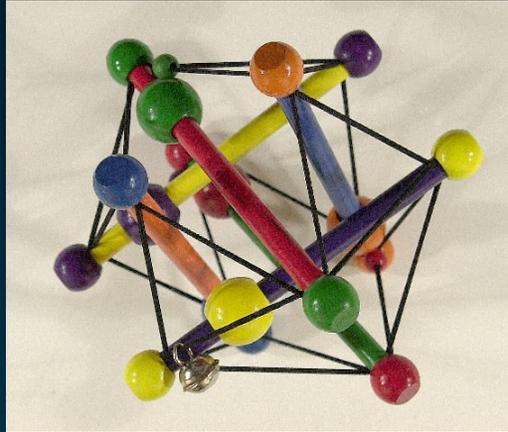
[Hahn & al., PSB, 2003]



Combine methods

- ◆ 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





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