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 (\textit{isa, part of})
    - Associative (\textit{location of, causes, ...})

◆ **Statistical relations**
  - Represented in text
    - Among lexical items (entity recognition)
    - Annotations
  - No explicit semantics
Example  Viral meningitis

CNS disease \(\text{isa}\) Infectious disease

Viral meningitis \(\text{isa}\) Meninges

Virus \(\text{caused by}\) Anitviral agents

T-Lymphocytes \(\text{treated by}\) 18

Herpesviridae Infections 9
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


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

CNS disease
Viral meningitis
aspirin
headache
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

**Continuants**

- **Dependent continuants**
  - require the existence of any other entity in order to exist? *yes*

- **Independent continuants**
  - require the existence of any other entity in order to exist? *no*

**Occurrents**

- unfold through time in successive phases

- Oxygen transport

- Glucose metabolism

- Echocardiography

- Oxygen transporter

- B-lymphocyte

- Mitral valve

- Streptococcus
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**
  - \(P(E_1 \cap E_2) = P(E_1) \cdot P(E_2)\)

- **Tests of independence**
  - \(\chi^2\) 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 ≥ 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

<table>
<thead>
<tr>
<th>Pathological process</th>
<th>Anatomical entity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Musculoskeletal Diseases</td>
<td>Musculoskeletal System</td>
</tr>
<tr>
<td>Digestive System Diseases</td>
<td>Digestive System</td>
</tr>
<tr>
<td>Stomatognathic Diseases</td>
<td>Stomatognathic System</td>
</tr>
<tr>
<td>Respiratory Tract Diseases</td>
<td>Respiratory System</td>
</tr>
<tr>
<td>Nervous System Diseases</td>
<td>Nervous System</td>
</tr>
<tr>
<td>Eye Diseases</td>
<td>Sense Organs (+)</td>
</tr>
<tr>
<td>Urological and Male Genital Diseases</td>
<td>Urogenital System</td>
</tr>
<tr>
<td>Female Genital Diseases and Pregnancy Complica-</td>
<td>Urogenital System</td>
</tr>
<tr>
<td>tions</td>
<td>Embryonic Structures</td>
</tr>
<tr>
<td>Cardiovascular Diseases</td>
<td>Cardiovascular System</td>
</tr>
<tr>
<td>Hemic and Lymphatic Diseases</td>
<td>Hemic and Immune Systems</td>
</tr>
<tr>
<td>Skin Diseases</td>
<td>Integumentary System</td>
</tr>
<tr>
<td>Endocrine Diseases</td>
<td>Endocrine System</td>
</tr>
</tbody>
</table>
## Identifying dependence relations

### Results

*has_agent*

<table>
<thead>
<tr>
<th>Pathological Process</th>
<th>Pathogen</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bacterial Infection and Mycoses</td>
<td>Bacteria</td>
</tr>
<tr>
<td></td>
<td>Fungi</td>
</tr>
<tr>
<td>Virus Diseases</td>
<td>Viruses</td>
</tr>
<tr>
<td>Parasitic Diseases</td>
<td>Animals (+)</td>
</tr>
</tbody>
</table>
Identifying statistical relations

- Aggregation at the level of top-level categories

- Viral meningitis (C02, C10)
  - (C02, A08) and (C10, A08)

- Meninges (A08)
Identifying statistical relations

- Contingency table

<table>
<thead>
<tr>
<th>Indexed with term $A$</th>
<th>Indexed with term $B$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>$n_{AB}$</td>
<td>$n_{AB}$</td>
</tr>
<tr>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>$n_{aB}$</td>
<td>$n_{ab}$</td>
</tr>
</tbody>
</table>

- 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

<table>
<thead>
<tr>
<th>Category</th>
<th>A01</th>
<th>A02</th>
<th>A03</th>
<th>A04</th>
<th>A05</th>
<th>A06</th>
<th>A07</th>
</tr>
</thead>
<tbody>
<tr>
<td>Body Regions</td>
<td>0.00</td>
<td>0.00</td>
<td>1.50</td>
<td>0.50</td>
<td>0.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>Musculoskeletal System</td>
<td>1.50</td>
<td>1.50</td>
<td>0.00</td>
<td>0.50</td>
<td>1.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>Digestive System</td>
<td>1.50</td>
<td>1.50</td>
<td>0.00</td>
<td>1.00</td>
<td>9.08</td>
<td>3.50</td>
<td>5.17</td>
</tr>
<tr>
<td>Respiratory System</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>1.00</td>
<td>0.00</td>
<td>0.00</td>
<td>2.50</td>
</tr>
<tr>
<td>Urogenital System</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>Endocrine System</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>Cardiovascular System</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
</tbody>
</table>
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* $\rightarrow$ *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 and can help maintain terminologies.
References

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

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]
## GO and annotation databases

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

<table>
<thead>
<tr>
<th>Gene</th>
<th>GO:0000793</th>
<th>IDA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Brca1</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Gene</th>
<th>GO:0003684</th>
<th>IDA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Brca1</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Gene</th>
<th>GO:0003723</th>
<th>ISS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Brca1</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Gene</th>
<th>GO:0004553</th>
<th>ISS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Brca1</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Gene</th>
<th>GO:0005515</th>
<th>ISS, TAS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Brca1</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Gene</th>
<th>GO:0005622</th>
<th>IDA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Brca1</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Gene</th>
<th>GO:0005634</th>
<th>IDA, ISS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Brca1</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Gene</th>
<th>...</th>
</tr>
</thead>
<tbody>
<tr>
<td>Brca1</td>
<td></td>
</tr>
</tbody>
</table>
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

- **Genes**: g₁, g₂, …, gₙ
- **GO terms**: t₁, t₂, …, tₙ

**Annotation database**

The diagram illustrates the relationship between genes and their associated GO terms. Each gene is represented by a row in the left table, and each GO term is represented by a column in the right table. The intersection of genes and GO terms indicates their association.
1. Similarity in the vector space model

**Genes**

<table>
<thead>
<tr>
<th></th>
<th>g₁</th>
<th>g₂</th>
<th>...</th>
<th>gₙ</th>
</tr>
</thead>
<tbody>
<tr>
<td>t₁</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>t₂</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>tₙ</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**GO terms**

<table>
<thead>
<tr>
<th>t₁</th>
<th>t₂</th>
<th>...</th>
<th>tₙ</th>
</tr>
</thead>
<tbody>
<tr>
<td>t₁</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>t₂</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>tₙ</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Similarity matrix**

\[ \text{Sim}(t_i, t_j) = \overrightarrow{t_i} \cdot \overrightarrow{t_j} \]
Analysis of co-occurring GO terms

GO terms

Genes

g_1 | g_2 | ... | g_n
| t_1 | t_2 | ... | t_n

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 |
| ... | ... |
Analysis of co-occurring GO terms

- Statistical analysis: test independence
  - Likelihood ratio test (G²)
  - Chi-square test (Pearson’s χ²)

- Example from GOA (22,720 annotations)
  - GO:0006955 [BP] Freq. = 588
  - GO:0008009 [MF] Freq. = 53

<table>
<thead>
<tr>
<th></th>
<th>present</th>
<th>absent</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>GO:0006955</td>
<td>46</td>
<td>542</td>
<td>588</td>
</tr>
<tr>
<td>GO:0008009</td>
<td>53</td>
<td>22,125</td>
<td>22,720</td>
</tr>
</tbody>
</table>

Total

G² = 298.7
p < 0.000
Association rule mining

<table>
<thead>
<tr>
<th>GO terms</th>
<th>t₁</th>
<th>t₂</th>
<th>...</th>
<th>tₙ</th>
</tr>
</thead>
<tbody>
<tr>
<td>g₁</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>g₂</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>gₙ</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- **apriori**
  - Rules: t₁ \( \Rightarrow \) t₂
  - Confidence: > .9
  - Support: .05

Annotation database

transaction
## Examples of associations

<table>
<thead>
<tr>
<th>Association</th>
<th>VSM</th>
<th>COC</th>
<th>ARM</th>
<th>LEX</th>
</tr>
</thead>
<tbody>
<tr>
<td>MF: <em>potassium channel activity</em></td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>BP: <em>potassium ion transport</em></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MF: <em>chemokine activity</em></td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BP: <em>immune response</em></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CC: <em>hemoglobin complex</em></td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BP: <em>oxygen transport</em></td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MF: <em>taste receptor activity</em></td>
<td></td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BP: <em>perception of taste</em></td>
<td></td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MF: <em>metal ion transporter activity</em></td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BP: <em>metal ion transport</em></td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CC: <em>transport vesicle</em></td>
<td></td>
<td></td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>BP: <em>transport</em></td>
<td></td>
<td></td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>CC: <em>gap junction</em></td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BP: <em>cell communication</em></td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
## Associations identified

<table>
<thead>
<tr>
<th></th>
<th>VSM</th>
<th>COC</th>
<th>ARM</th>
<th>LEX</th>
</tr>
</thead>
<tbody>
<tr>
<td>MF-CC</td>
<td>499</td>
<td>893</td>
<td>362</td>
<td>917</td>
</tr>
<tr>
<td>MF-BP</td>
<td>3057</td>
<td>1628</td>
<td>577</td>
<td>2523</td>
</tr>
<tr>
<td>CC-BP</td>
<td>760</td>
<td>1047</td>
<td>329</td>
<td>2053</td>
</tr>
<tr>
<td>Total</td>
<td>4316</td>
<td>3568</td>
<td>1268</td>
<td>5493</td>
</tr>
</tbody>
</table>

7665 by at least one approach
## Associations identified

### VSM

<table>
<thead>
<tr>
<th></th>
<th>VSM</th>
<th>COC</th>
<th>ARM</th>
<th>LEX</th>
</tr>
</thead>
<tbody>
<tr>
<td>MF-CC</td>
<td>499</td>
<td>893</td>
<td>362</td>
<td>917</td>
</tr>
<tr>
<td>MF-BP</td>
<td>3057</td>
<td>1628</td>
<td>577</td>
<td>2523</td>
</tr>
<tr>
<td>CC-BP</td>
<td>760</td>
<td>1047</td>
<td>329</td>
<td>2053</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>4316</td>
<td>3568</td>
<td>1268</td>
<td>5493</td>
</tr>
</tbody>
</table>

MF: ice binding

BP: response to freezing
### Associations identified

<table>
<thead>
<tr>
<th></th>
<th>VSM</th>
<th>COC</th>
<th>ARM</th>
<th>LEX</th>
</tr>
</thead>
<tbody>
<tr>
<td>MF-CC</td>
<td>499</td>
<td>893</td>
<td>362</td>
<td>917</td>
</tr>
<tr>
<td>MF-BP</td>
<td>3057</td>
<td>1628</td>
<td>577</td>
<td>2523</td>
</tr>
<tr>
<td>CC-BP</td>
<td>760</td>
<td>1047</td>
<td>329</td>
<td>2053</td>
</tr>
<tr>
<td>Total</td>
<td>4316</td>
<td>3568</td>
<td>1268</td>
<td>5493</td>
</tr>
</tbody>
</table>

**MF**: chromatin binding  
**CC**: nuclear chromatin
## Associations identified

### ARM

<table>
<thead>
<tr>
<th>MF-CC</th>
<th>VSM</th>
<th>COC</th>
<th>ARM</th>
<th>LEX</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>499</td>
<td>893</td>
<td>362</td>
<td>917</td>
</tr>
<tr>
<td>MF-BP</td>
<td>3057</td>
<td>1628</td>
<td>577</td>
<td>2523</td>
</tr>
<tr>
<td>CC-BP</td>
<td>760</td>
<td>1047</td>
<td>329</td>
<td>2053</td>
</tr>
<tr>
<td>Total</td>
<td>4316</td>
<td>3568</td>
<td>1268</td>
<td>5493</td>
</tr>
</tbody>
</table>

MF: carboxypeptidase A activity  
BP: peptolysis and peptidolysis
Limited overlap among approaches

- Lexical vs. non-lexical

- Among non-lexical

Venn diagram showing overlap among different approaches:

- VSM
- ARM
- COC

Numbers inside the circles and intersections:

- VSM: 3689
- ARM: 121
- COC: 2587
- Intersection of VSM and ARM: 305
- Intersection of VSM and COC: 453
- Intersection of ARM and COC: 201
- Total in VSM: 5493
- Total in ARM: 309
- Total in COC: 305
- Total in all: 7665
- Intersection of all: 230

Note: The numbers represent counts or measures of overlap between different approaches.
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

<table>
<thead>
<tr>
<th></th>
<th>Physical entity</th>
<th>Function</th>
<th>Process</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Organism</strong></td>
<td>Gross anatomy</td>
<td>Organism functions</td>
<td>Organism processes</td>
</tr>
<tr>
<td><strong>Cell</strong></td>
<td>Cellular components</td>
<td>Cellular functions</td>
<td>Cellular processes</td>
</tr>
<tr>
<td><strong>Molecule</strong></td>
<td>Molecules</td>
<td>Molecular functions</td>
<td>Molecular processes</td>
</tr>
</tbody>
</table>

[adapted from B. Smith]
GO and other domains (revisited)

<table>
<thead>
<tr>
<th>Resolution</th>
<th>Physical whole</th>
<th>Physical part</th>
<th>Function</th>
<th>Process</th>
</tr>
</thead>
<tbody>
<tr>
<td>Organism</td>
<td>Organism components</td>
<td>Organism functions</td>
<td>Organism processes</td>
<td></td>
</tr>
<tr>
<td>Cell</td>
<td>Cellular components</td>
<td>Cellular functions</td>
<td>Cellular processes</td>
<td></td>
</tr>
<tr>
<td>Molecule</td>
<td>Molecular components</td>
<td>Molecular functions</td>
<td>Molecular processes</td>
<td></td>
</tr>
</tbody>
</table>

[adapted from B. Smith]
# Biological ontologies (OBO)

<table>
<thead>
<tr>
<th>Domain</th>
<th>Prefix</th>
<th>Files</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cell type</td>
<td>CL</td>
<td>cell.obo</td>
</tr>
<tr>
<td>Chemical entities of biological interest</td>
<td>CHEBI</td>
<td>ontology.obo</td>
</tr>
<tr>
<td>Mus adult gross anatomy</td>
<td>MA</td>
<td>MA.ontology</td>
</tr>
<tr>
<td>Plant anatomy</td>
<td>PO</td>
<td>anatomy.ontology and anatomy.definition</td>
</tr>
<tr>
<td>NCBI organismal classification</td>
<td>taxon</td>
<td>taxonomy.dat</td>
</tr>
<tr>
<td>Human disease</td>
<td>DOID</td>
<td>DO 08 18 03.txt</td>
</tr>
<tr>
<td>Mouse pathology</td>
<td>MPATH</td>
<td>mouse pathology.ontology</td>
</tr>
<tr>
<td>PATO</td>
<td>PATO</td>
<td>attribute and value.obo</td>
</tr>
<tr>
<td>Physical-chemical methods and properties</td>
<td>FIX</td>
<td>fix.ontology</td>
</tr>
<tr>
<td>Physico-chemical process</td>
<td>REX</td>
<td>rex.obo</td>
</tr>
</tbody>
</table>
## GO and other domains (revisited)

<table>
<thead>
<tr>
<th>Resolution</th>
<th>Physical whole</th>
<th>Physical part</th>
<th>Function</th>
<th>Process</th>
</tr>
</thead>
<tbody>
<tr>
<td>Organism</td>
<td>Organism components</td>
<td>Organism functions</td>
<td>Organism processes</td>
<td></td>
</tr>
<tr>
<td>Cell</td>
<td>Cellular components</td>
<td>Cellular functions</td>
<td>Cellular processes</td>
<td></td>
</tr>
<tr>
<td>Molecule</td>
<td>Molecular components</td>
<td>Molecular functions</td>
<td>Molecular processes</td>
<td></td>
</tr>
</tbody>
</table>

[adapted from B. Smith]
Integrating biological ontologies

- Organism
- Biological processes
- Cellular components
- Other OBO ontologies
- Molecule
- Molecular functions
- Cell
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)
<table>
<thead>
<tr>
<th>Chemical</th>
<th>CHEBI:ID</th>
<th>Biological Process</th>
<th>GO:ID</th>
</tr>
</thead>
<tbody>
<tr>
<td>iron</td>
<td>[CHEBI:18248]</td>
<td>iron ion transport</td>
<td>[GO:0006826]</td>
</tr>
<tr>
<td></td>
<td></td>
<td>iron superoxide dismutase activity</td>
<td>[GO:0008382]</td>
</tr>
<tr>
<td></td>
<td></td>
<td>vanadium-iron nitrogenase complex</td>
<td>[GO:0016613]</td>
</tr>
<tr>
<td>uronic acid</td>
<td>[CHEBI:27252]</td>
<td>uronic acid metabolism</td>
<td>[GO:0006063]</td>
</tr>
<tr>
<td></td>
<td></td>
<td>uronic acid transporter activity</td>
<td>[GO:0015133]</td>
</tr>
<tr>
<td>carbon</td>
<td>[CHEBI:27594]</td>
<td>response to carbon dioxide</td>
<td>[GO:0010037]</td>
</tr>
<tr>
<td></td>
<td></td>
<td>carbon-carbon lyase activity</td>
<td>[GO:0016830]</td>
</tr>
</tbody>
</table>
Quantitative results

- 2,700 ChEBI entities (27%) identified in some GO term
- 9,431 GO terms (55%) include some ChEBI entity in their names
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

Contact: olivier@nlm.nih.gov
Web: mor.nlm.nih.gov

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