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
- Terminology / Knowledge
- Knowledge acquisition/capture
- Information extraction

[Jacquemin]  
TKE, TIA  
K-CAP
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**
- **Terms within a vocabulary**
  - Clinical vocabularies
  - Gene Ontology
- **Terms across vocabularies**
  - SNOMED / LOINC
  - GO / ChEBI
- **Lexicon / Terms**
  - Semantic lexicon

References:

- [Baud & al., AMIA, 1998]
- [McDonald & al., AMIA, 1999]
- [Ogren & al., PSB, 2004]
- [Mungall, CFG, 2004]
- [Dolin, JAMIA, 1998]
- [Burgun, SMBM, 2005]
- [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

GO terms

<table>
<thead>
<tr>
<th></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>

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>

Annotation database
1. **Similarity in the vector space model**

- **GO terms**
  - $t_1$, $t_2$, ..., $t_n$

- **Genes**
  - $g_1$, $g_2$, ..., $g_n$

- **Similarity matrix**
  - $\text{Sim}(t_i, t_j) = \vec{t}_i \cdot \vec{t}_j$

### GO terms

<table>
<thead>
<tr>
<th></th>
<th>$t_1$</th>
<th>$t_2$</th>
<th>...</th>
<th>$t_n$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$t_1$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$t_2$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$t_n$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Analysis of co-occurring GO terms

GO terms

<table>
<thead>
<tr>
<th>Genes</th>
<th>g_1</th>
<th>g_2</th>
<th>...</th>
<th>g_n</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>t_1</td>
<td>t_2</td>
<td>...</td>
<td>t_n</td>
</tr>
</tbody>
</table>

Annotation database

| t_2-t_7 | 1 |
| t_2-t_9 | 1 |
| t_7-t_9 | 2 |
| ...     |   |
Analysis of co-occurring GO terms

- Statistical analysis: test independence
  - Likelihood ratio test ($G^2$)
  - Chi-square test (Pearson’s $\chi^2$)

- Example from GOA (22,720 annotations)

```markdown
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>GO:0008009 [MF]</td>
<td>C0008009 [MF]</td>
<td>Freq. = 53</td>
</tr>
</tbody>
</table>
```

Co-oc. = 46

<table>
<thead>
<tr>
<th>GO:0008009</th>
<th>immune response</th>
</tr>
</thead>
<tbody>
<tr>
<td>present</td>
<td>absent</td>
</tr>
<tr>
<td>present</td>
<td>46</td>
</tr>
<tr>
<td>absent</td>
<td>7</td>
</tr>
<tr>
<td>total</td>
<td>53</td>
</tr>
</tbody>
</table>

$G^2 = 298.7$

p < 0.000
Association rule mining

GO terms

<table>
<thead>
<tr>
<th>Genes</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>

annotation database

.transaction

g₂

t₁ \Rightarrow t₂

Rules: t₁ \Rightarrow t₂
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