Slicing through the Scientific Literature

Christopher Baker (1), Patrick Lambrix (2), Jonas Laurila Bergman (2), Rajaraman Kanagasabai (3), Wee Tiong Ang (3)

(1) University of New Brunswick, (2) Linköpings Universitet, (3) ASTAR Singapore
Literature search

- Huge amount of scientific literature.
- Need to integrate a spectrum of information to perform a task.
Literature search

- How to know what is in the repository
  - Lack of knowledge of the domain

- How to compose an expressive query
  - Lack of knowledge of search technology
Example scenario

“Lipid”

- Keyword search returns all documents containing lipid.
  - No knowledge; terminology problem
- Relationships: use of multiple keywords with/without boolean operators,
  e.g. *lipid and disease*
Example scenario

“Lipid”

- Keyword search returns a list of relevant questions concerning lipid. User selects question and retrieves knowledge and provenance documents.

- Multiple search terms: requirement that there are relevant connections between the keywords.
lipid
KnowleFinder

1. Which lipid has a broad synonym
2. Which lipid has a lipid KEGG ID and has a broad synonym
3. Which lipid is implicated in a disease
4. Which lipid interacts with proteins
5. Which lipid is implicated in a disease and interacts with proteins
6. Which lipid is implicated in a disease and interacts with proteins involved in signal pathways
7. Which lipid is found in a sentence is implicated in a disease and interacts with proteins involved
8. Which document contains a sentence in which lipid is implicated in a disease and interacts with proteins involved
Question
NLG: Which lipid is implicated in a disease and interacts with proteins involved in signal pathways?

Result

<table>
<thead>
<tr>
<th>Protein</th>
<th>Lipid</th>
<th>Disease</th>
<th>Signal Pathway</th>
</tr>
</thead>
<tbody>
<tr>
<td>P53</td>
<td>Unsat. Fatty Acid</td>
<td>Ovarian Cancer</td>
<td>Apoptosis</td>
</tr>
</tbody>
</table>
Example scenario

Requirements

- Natural language interface
- Ontology-driven query
- Context of query terms
- Cross-domain queries
Outline

- Relevant queries
- Framework for slicing through the scientific literature
- Algorithms and example
- Conclusion & Future Work
Outline

- Relevant queries
- Framework for slicing through the scientific literature
- Algorithms and example
- Conclusion & Future Work
Ontologies define the basic terms and relations comprising the vocabulary of a topic area, as well as the rules for combining terms and relations to define extensions to the vocabulary.”
### Ontologies

**GENE ONTOLOGY (GO)**

<table>
<thead>
<tr>
<th>Concept</th>
</tr>
</thead>
<tbody>
<tr>
<td>immune response</td>
</tr>
<tr>
<td>- acute-phase response</td>
</tr>
<tr>
<td>- anaphylaxis</td>
</tr>
<tr>
<td>- antigen presentation</td>
</tr>
<tr>
<td>- antigen processing</td>
</tr>
<tr>
<td>- cellular defense response</td>
</tr>
<tr>
<td>- cytokine metabolism</td>
</tr>
<tr>
<td>- cytokine biosynthesis</td>
</tr>
<tr>
<td>synonym cytokine production</td>
</tr>
<tr>
<td>- regulation of cytokine biosynthesis</td>
</tr>
<tr>
<td>- B-cell activation</td>
</tr>
<tr>
<td>- B-cell differentiation</td>
</tr>
<tr>
<td>- B-cell proliferation</td>
</tr>
<tr>
<td>- cellular defense response</td>
</tr>
<tr>
<td>- T-cell activation</td>
</tr>
<tr>
<td>- activation of natural killer</td>
</tr>
<tr>
<td>cell activity</td>
</tr>
</tbody>
</table>

We will assume a graph representation.
Relevant queries

n Relevant query including a number of concepts and relations from an ontology

connected sub-graph of the ontology that includes the concepts and relations.

(query graph based on the concepts and relations; slice is set of all query graphs based on the concepts and relations)
Query graph
Query graph
Query graph
Special cases

- No relations, several concepts
  - Relevant queries regarding concepts; relations are suggested by the system.
  - Difference with traditional techniques: extra requirement that search terms need to be connected in the ontology.

- No relations, one concept
  - Relevant queries including a specific query term.
  - Computes the ontological environment of the query term.
Multiple ontologies

<table>
<thead>
<tr>
<th>GENE ONTOLOGY (GO)</th>
<th>SIGNAL-ONTOLOGY (SigO)</th>
</tr>
</thead>
<tbody>
<tr>
<td>immune response</td>
<td>Immune Response</td>
</tr>
<tr>
<td>i- acute-phase response</td>
<td>i- Allergic Response</td>
</tr>
<tr>
<td>i- anaphylaxis</td>
<td>i- Antigen Processing and Presentation</td>
</tr>
<tr>
<td>i- antigen presentation</td>
<td>i- B Cell Activation</td>
</tr>
<tr>
<td>i- antigen processing</td>
<td>i- B Cell Development</td>
</tr>
<tr>
<td>i- cellular defense response</td>
<td>i- Complement Signaling</td>
</tr>
<tr>
<td>i- cytokine metabolism</td>
<td>synonym complement activation</td>
</tr>
<tr>
<td>i- cytokine biosynthesis</td>
<td>i- Cytokine Response</td>
</tr>
<tr>
<td>synonym cytokine production</td>
<td>i- Immune Suppression</td>
</tr>
<tr>
<td>…</td>
<td>i- Inflammation</td>
</tr>
<tr>
<td>…</td>
<td>i- Intestinal Immunity</td>
</tr>
<tr>
<td>…</td>
<td>i- Leukotriene Response</td>
</tr>
<tr>
<td>…</td>
<td>i- Leukotriene Metabolism</td>
</tr>
<tr>
<td>…</td>
<td>i- Natural Killer Cell Response</td>
</tr>
<tr>
<td>…</td>
<td>i- T Cell Activation</td>
</tr>
<tr>
<td>…</td>
<td>i- T Cell Development</td>
</tr>
<tr>
<td>…</td>
<td>i- T Cell Selection in Thymus</td>
</tr>
<tr>
<td>…</td>
<td></td>
</tr>
</tbody>
</table>
Ontology Alignment

Alignment is a set of mappings between terms in the ontologies.
Relevant queries – multiple ontologies

Relevant query including a number of concepts and relations from multiple ontologies

Query graphs connected by a path going through a mapping in the alignment.

(aligned query graph based on query graphs; aligned slice is set of all aligned query graphs based on the query graphs)
Aligned query graph
Aligned query graph
Aligned query graph
Outline

- Relevant queries
- **Framework for slicing through the scientific literature**
- Algorithms and example
- Conclusion & Future Work
Framework
Framework

- External resources
  - Literature document base
  - Ontology and ontology alignment repository

- Computed resources
  - Knowledge base
Framework

- Process and translation
  - Knowledge base instantiation
  - Slice generation and alignment
  - Natural language query generation

- Query
Outline

- Relevant queries
- Framework for slicing through the scientific literature
- Algorithms and example
- Conclusion & Future Work
External resources

- Literature document base
  - Generated from a collection of 7498 PubMed abstracts relevant for Ovarian Cancer. 683 papers included lipid names from which 241 full papers were downloadable.

- Ontology and ontology alignment repository
  - Lipid ontology
  - Signal ontology
  - Alignment using SAMBO
Lipid research is increasingly integrated within systems level biology such as lipidomics where lipid classification is required before appropriate annotation of chemical functions can be applied. The ontology describes the LIPIDMAPS nomenclature classification explicitly using description logics (OWL-DL). Lipid classes are organized hierarchically with the super-classes restricted by generic necessary conditions. More specific necessary conditions are used to define membership requirements for sub-classes of lipid according to appropriate functional groups.
SAMBO (1)

- **SAMBO** (System for Aligning and Merging Biomedical Ontologies)
  - Phase I
    - Matchers
    - Weighted sum combination of matcher results
    - Single threshold filtering

*Winner Anatomy Track of OAEI 2008*
SAMBO (2)

- Phase II:

![Diagram showing the process of suggestions and conflict checker]
Knowledge base instantiation

BioText toolkit

- Named entity recognition
  - Gazetteer approach
  - Termlists
    - Lipids: LIPIDMAPS, LipidBank, KEGG, IUPAC
    - Proteins: Swiss-Prot
    - Diseases: disease ontology of Center of Genetic Medicine
  - Tags found entities with ontology concepts

- Normalization and grounding

- Relation detection
  - Based on co-occurrence in sentence
  - + rule set developed with domain expert
Knowledge base instantiation

1) Document Content
2) Sentence Extraction
3) Sentence Detection: lipid interaction protein
4) Entity Recognition: term identification / assign lipid class
5) Normalization: collapse lipid synonyms
6) Relation Extraction: Lipid-Protein or Lipid Disease
   "TLR4 binds to POPC", tagged as
   "<term category="protein"> TLR4</term> binds to
   <term category="lipid">POPC</term>"
7) Classification: Identify ontology classes and specify relations for all sentences, proteins, lipid subclasses.
8) Populate OWL ontology (JENA -API)
Mined Interactions

Table 1. Interactions mined from the ovarian cancer bibliome. OC and AP represent a cancer and apoptosis pathway proteins respectively.

<table>
<thead>
<tr>
<th>Interaction Type</th>
<th>Abstract (7498)</th>
<th>Full Paper (241)</th>
</tr>
</thead>
<tbody>
<tr>
<td>OC-OC</td>
<td>223</td>
<td>13</td>
</tr>
<tr>
<td>OC-AP</td>
<td>505</td>
<td>195</td>
</tr>
<tr>
<td>OC-Lipid</td>
<td>11</td>
<td>14</td>
</tr>
<tr>
<td>OC-Hormone</td>
<td>8</td>
<td>1</td>
</tr>
<tr>
<td>AP-AP</td>
<td>113</td>
<td>59</td>
</tr>
<tr>
<td>AP-Lipid</td>
<td>10</td>
<td>8</td>
</tr>
<tr>
<td>Lipid-Lipid</td>
<td>3</td>
<td>23</td>
</tr>
<tr>
<td>Lipid Hormone</td>
<td>2</td>
<td>18</td>
</tr>
<tr>
<td>Protein Hormone</td>
<td>9</td>
<td>2</td>
</tr>
<tr>
<td>Hormone-Hormone</td>
<td>2</td>
<td>6</td>
</tr>
</tbody>
</table>
Knowledge base instantiation

Population

Concepts

Properties

Knowledge Provenance documents
Knowledge base instantiation
Slice generation

- Current implementation focuses on slices based on concepts.
- Depth-first traversal of ontology to find paths between given concepts; paths can be put together to find slices/query graphs.
Slice alignment

- Algorithm computes subset of aligned slice.
- Assumption: shorter paths represent closer relationships.
- Algorithm connects slices using shortest paths from given concepts in one ontology to given concepts in other ontology.
Slicing through the literature

nRQL: (RETRIEVE (?X ?Y ?Z ?W)
(AND (?X Protein) (?Y Lipid) (?Z Disease) (?W SignalPathway)
(?X ?Y Interacts with) (?Y ?Z Implicated in) (?X ?W Involved in)))
Natural language query generation

- Triple representation:
  \[ \langle \text{lipid}, \text{interacts-with}, \text{protein} \rangle \]

- Rule base to generate NL statements.
  
  \textit{What lipid interacts with proteins?}

  - Learned from examples.

- Aggregation of statements from different triples, grammar checking.
KnowleFinder

1. Which lipid has a broad synonym
2. Which lipid has a lipid KEGG ID and has a broad synonym
3. Which lipid is implicated in a disease
4. Which lipid interacts with proteins
5. Which lipid is implicated in a disease and interacts with proteins
6. Which lipid is implicated in a disease and interacts with proteins involved in signal pathways
7. Which lipid is found in a sentence is implicated in a disease and interacts with proteins involved
8. Which document contains a sentence in which lipid is implicated in a disease and interacts with proteins involved
n Send nRQL query to RACER.

Question
NLG: Which lipid is implicated in a disease and interacts with proteins involved in signal pathways?

nRQL: (RETRIEVE (?X ?Y ?Z ?W)
(AND (?X Protein) (?Y Lipid) (?Z Disease) (?W SignalPathway)
(?X ?Y Interacts_with) (?Y ?Z Implicated_in) (?X ?W Involved_in)))

Result

<table>
<thead>
<tr>
<th>Protein</th>
<th>Lipid</th>
<th>Disease</th>
<th>Signal Pathway</th>
</tr>
</thead>
<tbody>
<tr>
<td>P53</td>
<td>Unsat. Fatty Acid</td>
<td>Ovarian Cancer</td>
<td>Apoptosis</td>
</tr>
</tbody>
</table>
Outline

- Relevant queries
- Framework for slicing through the scientific literature
- Algorithms and example
- **Conclusion & Future Work**
Conclusions

- Framework for literature search dealing with
  - Lack of knowledge of the domain.
  - Lack of knowledge of the search technology.

- Proof of concept implementation.
Future Work

- Heuristics and their influences for slice generation.
- Tradeoff in query generation between completeness and information overload.
- Relevance measure and query ranking.
- Optimization of slice generation.
Future Work

- Integration in larger system
- Integrated implementation.
- Scalability testing.