System Description: Reaching Deeper into the Life Science Bibliome with CORAAL

Vít Nováček    Tudor Groza    Siegfried Handschuh

Digital Enterprise Research Institute (DERI)
National University of Ireland Galway
IDA Business Park, Lower Dangan, Galway, Ireland

e-mail: FirstName.LastName@deri.org

SWASD @ ISWC 2009, Chantilly, VA, USA

October 26, 2009
Outline

1. Introduction
   - Problem Formulation
   - Proposed Solution

2. Method
   - Processed Data
   - CORAAL Work-Flow

3. Usage Examples
   - Asking Queries
   - Browsing Query Results
   - Filtering Classical Search

4. Conclusion
   - Conclusions and Future Work
   - Additional Information
Motivating Use Case

**exploiting knowledge in biomedical resources**
- analysis of semi-structured electronic patient records
  - identification of entities occurring in particular fields
  - finding simple relationships between them
- extraction of expressive statements from articles

**translational biomedical knowledge integration**
- integrating the extracted content
  - identifying related patient records
  - linking the patient and publication knowledge
  - refinement of the integrated knowledge
- intuitive querying of the integrated content
Can We Search for Knowledge Already?

- Search engines and text mining are good...
  - Knowledge more open and exploitable
  - Enabling efficient search for terms
  - Retrieval and browsing of additional annotations

- But still have huge reserves
  - Floating on the surface of raw text and meta-data
  - Reaching deeper for the knowledge is difficult
    - Requires expressive models
    - Expensive and tedious to acquire
    - Neither scalable, nor economical
  - Text mining solutions provide automation
  - Rather shallow insights and limited querying capabilities
A Way to Go

let the machines work

- extract automatically:
  - annotations of publication content and meta-data
  - knowledge from publications
    - the emergent knowledge is rather noisy and sparse, though
    - further processing needed

tackle the emergent knowledge

- integration with precise domain resources
- expansion and refinement by means of inference
- intuitive expressive querying
- semi-automatic refinement using user feedback
Current CORAAL Repository

- 12,000 articles
- 7,500,000 meta-data
- 15,000,000 knowledge
## Content Description

### article (meta)data
- source – articles
- text and respective full-text indices
- metadata:
  - titles, author names, references, citation contexts
  - stored as triple annotations in an RDF repository

### publication knowledge
- source – articles and NCI, Emtree thesauri
- concepts and their synonyms
- uncertain argument-relationship-argument statements in context
- types of relationships:
  - taxonomical (type)
  - difference (negative type)
  - generic facets (e.g., involves expression of, part of)
example article (meta)data

- **title**: *On Oncological Hematologic Disorders*
- **author**: John Smith
- **text**: ... *The rate of T-cell leukemia, acute granulocytic leukemia and other hematologic disorders in the studied sample was about three times higher than average. ...*
- **references, referenced-by, etc.**

extracted content

- respective meta-data as a **machine-readable graph**
- **quad statements** representing extracted knowledge:
  - *(T-cell leukemia:is a:leukemia, 1.0)*
  - *(T-cell leukemia:is a:acute granulocytic leukemia, -0.6)*
  - *(T-cell leukemia:is a:hematologic disorder, -0.6)*
Processing the Extracted Information

integration
- incrementally merging and interlinking the extracted content

refinement
- a mistake – *T-cell leukemia* actually is a *hematologic disorder*
- according to the NCI thesaurus, we can infer that:
  $(\text{leukemia}: \text{is a}: \text{hematologic disorder}, 1.0)$
- in effect, mistake resolved to:
  $(\text{T-cell leukemia}: \text{is a}: \text{hematologic disorder}, 0.73)$

extension
- extracted concepts linked to analogical knowledge
- examples: interleukin 6 $\leftrightarrow$ antigen, transcription initiation site $\leftrightarrow$ introns...
Diving into the Content

**publication data and meta-data**
- classical **full-text search** in publication *texts, titles or authors*
- rather standard **key-word queries**
- result **filtering** using automatically associated **concepts**

**publication knowledge**
- search for **knowledge** related to mere **terms**
- search using an **expressive**, yet simple **query language**
- examples:
  - cystosarcoma phylloides
  - acute granulocytic leukemia : NOT is a : T-cell leukemia
  - ? : used as : complementary method AND ? :
    NOT is a : polymerase chain reaction
Outline

1 Introduction
   - Problem Formulation
   - Proposed Solution

2 Method
   - Processed Data
   - CORAAL Work-Flow

3 Usage Examples
   - Asking Queries
   - Browsing Query Results
   - Filtering Classical Search

4 Conclusion
   - Conclusions and Future Work
   - Additional Information
Query Construction

Knowledge query builder

NOT Subject: acute granulocytic leukemia

Relation: is a

Object: T-cell leukemia

Query: 

Submit
Introduction

Method

Usage Examples

Conclusion

Asking Queries

Browsing Query Results

Filtering Classical Search

Query Result

**acute granulocytic leukemia NOT TYPE T-cell leukemia**

**Sources:**

- Coding sequence and intron\xe2\x80\x93exon junctions of the c-myb gene are intact in the chron...

---

**Title:** Coding sequence and intron\xe2\x80\x93exon junctions of the c-myb gene are intact in the chronic phase and blast crisis stages of chronic myeloid leukemia patients

**Authors:** D Colomer, B. Calabretta, C. Silvestri, G. Martinelli, F. Cervantes, F. Bussolari, O. Candini, F. Corradini, C. Guerzoni, S.A. Mariani, S. Cattelani, L. Pecorari, I. Iacobucci, S. Soverini, T. Fasano

**Abstract:** The c-myb gene encodes a transcription factor required for proliferation, differentiation and survival of normal and leukemic hematopoietic cells. c-Myb has a longer half-life in BCR/ABL-expressing than in normal cells, a feature which depends, in part, on PI-3K/Akt-dependent regulation of proteins interacting with the leucine zipper/negative regulatory region of c-Myb. Thus, we asked whether the stability of c-Myb in leukemic cells might be enhanced by mutations interfering with its degradation. We analyzed the c-myb gene in 133 chronic myeloid leukemia (CML) patients in chronic phase and/or blast crisis by denaturing-high performance liquid chromatography (D-HPLC) and sequence analysis of PCR products corresponding to the entire coding sequence and each exon?intron boundary. No mutations were found. We found four single nucleotide polymorphisms (SNPs) and identified an alternatively spliced transcript lacking exon 5, but SNPs frequency and expression of the alternatively spliced transcript were identical in normal and CML cells. Thus, the enhanced stability of c-Myb in CML blast crisis cells and perhaps in other types of leukemia is not caused by a genetic mechanism.

**Certainty:** 0.6640

**Contexts:** oncology, genetics, pharmacology, biochemistry, biology, cell_research, and clinical_medicine

**Inferred:** false
Results of ? : is a : breast cancer
### Focusing on Particular Results

<table>
<thead>
<tr>
<th>Query</th>
<th>Result</th>
</tr>
</thead>
<tbody>
<tr>
<td>breast carcinoma NOT TYPE lung cancer</td>
<td><img src="image1" alt="Result" /></td>
</tr>
<tr>
<td>breast carcinoma HAS PART epigenetic silencing</td>
<td><img src="image2" alt="Result" /></td>
</tr>
</tbody>
</table>
Focusing on Authors Writing about Particular Topics

Name: A Lin
Affiliation: School of Medicine, Taipei Medical University, Taipei, Taiwan, ROC
Address: Not available.
Publications:
- Casein kinase II is a negative regulator of c-Jun DNA binding and AP-1 activity
- Protein damage and degradation by oxygen radicals. II. Modification of amino acids
- Coordinate regulation of I?B kinases by mitogen-activated protein kinase 1 and NF-?B-inducing kinase
- Differential activation of ERK and JNK mitogen-activated protein kinases by ref-1 and MERK

Focus on (?)

Direct concepts (?)
1. academic research enhancement awards
2. correlated
3. incubator

Super concepts (?)
1. gene amplification abnormality
2. myeloid
3. nucleic acid hybridization

V. Nováček, T. Groza, S. Handschuh
Conclusions and Future Work

**Conclusion**

- CORAAL presents a **first step** towards realising the **use case**
- continuous **tests** and **interviews** with experts show we are on the **right path**
- Elsevier Grand Challenge **results** seem to mean something, too ;)

**Future Work**

- two **major issues** remain to be tackled:
  - i. **wisdom of the crowds** exploitation
  - ii. move from CORAAL to a CORAAL **reef**
## Additional Information

<table>
<thead>
<tr>
<th>CORAAL web interface</th>
<th><a href="http://coraal.deri.ie:8080/coraal/">http://coraal.deri.ie:8080/coraal/</a></th>
</tr>
</thead>
<tbody>
<tr>
<td>CORAAL quick-start</td>
<td><a href="http://smile.deri.ie/projects/egc/quickstart">http://smile.deri.ie/projects/egc/quickstart</a></td>
</tr>
<tr>
<td>CORAAL demo</td>
<td><a href="http://resources.smile.deri.ie/coraal/videos/coraal_web.mp4">http://resources.smile.deri.ie/coraal/videos/coraal_web.mp4</a></td>
</tr>
</tbody>
</table>