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

CORAAL
dive into publications, bathe in the knowledge

SmILE Research Group
DERI

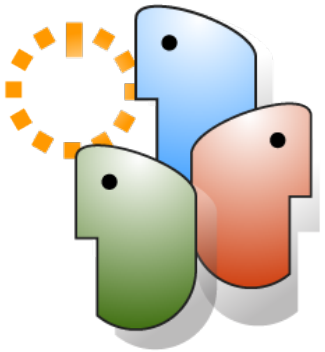
Outline

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 then
- 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 textmining 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

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

Current CORAAL Repository

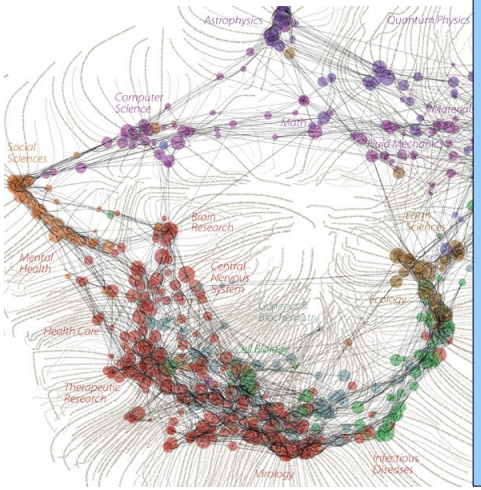
12,000

articles

7,500,000

meta-data

15,000,000

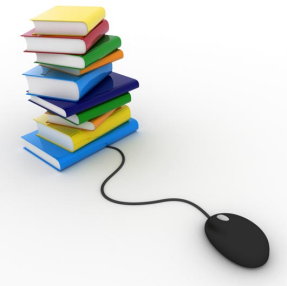


knowledge

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*)

Feeding on the Texts

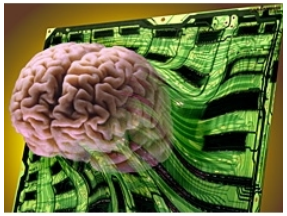
example article (meta)data

- 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:
  `(leukemia:is a:hematologic disorder, 1.0)`
- in effect, mistake resolved to:
  `(T-cell leukemia:is a:hematologic disorder, 0.73)`

extension

- extracted concepts linked to analogical knowledge
- examples: `interleukin 6` ↔ `antigen`, `transcription initiation site` ↔ `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

CORAAL

dive into publications, bathe in the knowledge

Knowledge query builder

NOT Subject: acute granulocytic leukemia

Relation: is a

Object: T-cell leukemia

T-cell leukemia

AND

Query:

Submit

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?exon 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, R. 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*

Subject (?)

- 61 breast carcinogenesis
- 61 breast carcinoma
- 29 breast metastasis
- 26 cystosarcoma phylloides
- 24 male breast carcinoma
- 20 breast sarcoma
- 19 breast adenocarcinoma
- 19 female breast carcinoma
- 19 intraductal carcinoma

Property (?)

- 275 TYPE
- 12 LINK
- 7 HAS PART
- 2 PART OF
- 1 PART

Object (?)

- 11 neoplasm
- 10 breast cancer
- 10 breast disease
- 10 breast tumor
- 10 disease, disorder or finding
- 10 disease or disorder
- 10 disorder by site
- 10 eo_findings_and_disorders_kind

297 Statement

sorted by: rank; then by... • grouped as sorted

breast sarcoma **TYPE** **breast cancer**
Sources:
Certainty: 1.0
Contexts: biochemistry, oncology, cell_research, pharmacology, and clinical_medicine
Inferred: true

Paget nipple disease **TYPE** **breast cancer**
Sources:
Certainty: 1.0
Contexts: biochemistry, oncology, cell_research, pharmacology, and clinical_medicine
Inferred: true

breast carcinogenesis **TYPE** **breast cancer**
Sources:
Certainty: 1.0
Contexts: biochemistry, oncology, cell_research, pharmacology, and clinical_medicine
Inferred: true

breast metastasis **TYPE** **breast cancer**
Sources:
Certainty: 1.0
Contexts: biochemistry, oncology, cell_research, pharmacology, and clinical_medicine
Inferred: true

breast carcinoma **TYPE** **breast cancer**
Sources:
- Pulmonary delivery of drugs for bone disorders
- Metallothionein immunostaining in adenoid cystic carcinomas of the salivary glands
- Prognostic significance of cyclin D1 amplification and overexpression in oral squamou...
- More_sources

Certainty: 1.0
Contexts: biochemistry, oncology, cell_research, pharmacology, and clinical_medicine

Focus on (?)

Context (?)

biochemistry biology
cell_research
clinical_medicine genetics
oncology pharmacology

Sources (?)

- 21 Comparative genomic hybridization analysis of benign and invasive male breast neoplas...
- 16 Pulmonary delivery of drugs for bone disorders
- 10 Metallothionein immunostaining in adenoid cystic carcinomas of the salivary glands
- 9 Application of comparative genomic hybridization in search for genetic aberrations in...
- 9 Neoplasms associated with paraneoplastic pemphigus: a review with emphasis on non-hem...
- 8 Allelic imbalance of HER2 variant in

Inferred statements (?)

- 274 true
- 23 false

Negative statements (?)

- 266 false
- 31 true

Focusing on Particular Results

3 Statement filtered from 297 originally (Reset All Filters)

sorted by: rank; then by... • grouped as sorted

breast carcinoma NOT TYPE lung cancer

Sources:

- Constitutional t(3;11)(p21;q23) in a Family, Including One Member with Lymphoma
- Down-regulation of Cdk inhibitor p27 in oral squamous cell carcinoma
- Hematopoietic Growth Factors in Oncology: Basic Science and Clinical Therapeutics
- More_sources

Certainty: 0.7980
Contexts: cell_research
Inferred: false

breast carcinoma HAS PART epigenetic silencing

Sources:

- The ATM/p53 pathway is commonly targeted for inactivation in squamous cell carcinoma ...
- Effects of demethylating agent 5-aza-2\xe2\x80\xb2-deoxycytidine and histone deacetylase inhib...
- DRAM, a p53-Induced Modulator of Autophagy, Is Critical for Apoptosis

Certainty: 0.8000
Contexts: oncology, genetics, pharmacology, biochemistry, biology, cell_research, and clinical_medicine
Inferred: true

Focusing on Authors Writing about Particular Topics

2 Author filtered from 50 originally (Reset All Filters)

sorted by: hasRank; then by... • grouped as sorted

Name: A Lin
Details ...

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 raf-1 and MEKK

Focus on (?)

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

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

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

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

CORAAL web interface

- http://coraal.deri.ie:8080/coraal/

CORAAL quick-start

- http://smile.deri.ie/projects/egc/quickstart

CORAAL demo

- http://resources.smile.deri.ie/coraal/videos/coraal_web.mp4

Emergent Knowledge Processing Details

- http://140.203.154.209/~vit/resources/2009/pubs/iswc2009n.pdf