Annotating the Biomedical Literature through Text Mining

Karin Verspoor, Ph.D.
Faculty, Computational Bioscience Program
University of Colorado School of Medicine

Karin.Verspoor@ucdenver.edu
http://compbio.ucdenver.edu/Hunter_lab/Verspoor
Exponential knowledge growth

- 1,330 peer-reviewed gene-related databases in 2011 NAR db issue
- Over 20 million PubMed entries (> 2,200/day)
- Breakdown of disciplinary boundaries makes more of it relevant to each of us
- “Like drinking from a firehose” – Jim Ostell
The motivation for text mining

1,000 genomes project will create 1,400GB this year
http://1000genomes.org

\begin{align*}
y &= \sim e^{0.031x} \\
R^2 &= 0.95
\end{align*}

\begin{align*}
y &= \sim e^{0.0418x} \\
R^2 &= 0.99
\end{align*}
BIONLP APPLICATIONS
Search is the most obvious application

- From keyword search

- To semantic search
  - Concepts (e.g. index by gene *identifiers*)
  - Relations (e.g. protein interactions)
1. Select papers from MEDLINE
2. List genes for curation
3. Curate genes from paper
4. Record associations in db

From Hirschman et al. BMC Bioinformatics 2005 6(Suppl 1):S1
The significance of the interaction between DAZAP1 and DAZL/DAZ remains to be defined. These proteins may act together to facilitate the expression of a set of genes in germ cells. For example, DAZAP1 could be involved in the transport of the mRNAs of the target genes of DAZL. Alternatively, DAZL and DAZAP1 may act antagonistically to regulate the timing and the level of expression. Such an antagonistic interaction between two interacting RNA-binding proteins is exemplified by the neuron-specific nuclear RNA-binding protein, Nova-1. Nova-1 regulates the alternative splicing of the pre-mRNAs encoding neuronal inhibitory glycine receptor α2 (GlyR α2) [23]. The ability of Nova-1 to activate exon selection in neurons is antagonized by a second RNA-binding protein, brPTB (brain-enriched polypyrimidine tract-binding protein), which interacts with Nova-1 and inhibits its function [24]. DAZAP1 could function in a similar manner by binding to DAZL and inhibiting its function. Comparing the phenotypes of Dazl1 and Dazap1 single and double knock-out mice may provide some clues to the significance of their interaction. Dazl1 knock-out mice have already been generated and studied [6]. The spermatogenic defect in the male becomes apparent only after day 7 post partum when the germ cells are committing to meiosis (H. Cooke, personal communication). The genomic structure of Dazap1, delineated here, should facilitate the generating of Dazap1 null mutation.
Scientific Publishing & Semantics

• Content enrichment
• Direct access to (relevant) external data
• Structured digital abstracts

• Enables
  – Interactivity
  – targeted searches
  – relevance linking
  – formalizing content; actionable data
Community-oriented Information Extraction

• Community ontologies:
  – We are committed to taking advantage of resources in the Open Biomedical Ontologies Foundry
  – We currently work primarily with Gene Ontology (GO), Chemical Entities of Biological Interest (ChEBI), and Cell Ontology

• Community platform:
  – Our system is implemented in Apache UIMA (Unstructured Information Management Architecture)
  – Several groups in BioNLP use UIMA and we are trying to achieve consensus on a shared type system
KaBOB (Knowledge Base Of Biology)

- Sources of Biological Data:
  - DIP
  - Entrez Gene
  - GAD
  - GOA
  - HGNC
  - HomoloGene
  - HPRD
  - InterPro
  - iRefWeb
  - KEGG
  - MGI
  - OMIM
  - PharmGKB
  - PReMod
  - Reactome
  - TRANSFAC
  - UniProt
  - ...

- Unify into a single knowledge base, preserving provenance from the existing data sources, and standardizing identifiers

- Data represented in RDF/OWL (semantic web)

- 900 million assertions (RDF triples) and...
Biological Knowledge Discovery

INFORMATION EXTRACTION
OpenDMAP extracts typed relations from the literature

- Concept recognition tool
  - Connect ontological terms to literature instances
  - Built on Protégé knowledge representation system

- Language patterns associated with concepts and slots
  - Patterns can contain text literals, other concepts, constraints (conceptual or syntactic), ordering information, or outputs of other processing.
  - Linked to many text analysis engines via UIMA

- Best performance in BioCreative II IPS task; high recall in BioNLP’09

OpenDMAP and annotations

- OpenDMAP is both a consumer and a producer of annotations

**Consumer:**
- Patterns defined in terms of semantic categories (concepts previously recognized)
- Semantic constraints defined in an ontological resource

**Producer:**
- Recognizes concepts and relations from text
- Explicitly represented as annotations over the text
OpenDMAP

Direct Memory Access Parser

- Shared “Memory”
- Uses ontological constraints in pattern matching
- Semantic grammar, combining text literals, syntactic categories and semantic constraints
Cyclin E2 interacts with Cdk2 in a functional kinase complex.

<ontology>
Protein protein interaction := [int1] interacts with [int2]

protein protein interaction:
interactor1: cyclin E2
interactor2: cdk2
OpenDMAP

**PROTÉGÉ ONTOLOGY**

**CLASS**: protein protein interaction  
**SLOT**: interactor1  
  TYPE: molecule  
**SLOT**: interactor2  
  TYPE: molecule

**PATTERNS**

\{c-interact\} := [interactor1] interacts with [interactor2]  
\{c-interact\} := [interactor1] is bound by [interactor2]  
...
Some BioCreative patterns for interact

\{c-interact\} := [\text{interactor}_1] \{\text{w-is}\} \{\text{w-interact-verb}_1\} \{\text{w-preposition}\} \text{ the?} \\
[\text{interactor}_2];

\{\text{w-is}\} := \text{is, are, was, were};

\{\text{w-interact-verb}_1\} := \text{co-immunoprecipitate, co-immunoprecipitates, co-immunoprecipitated, co-localize, co-localizes, co-localized};

\{\text{w-preposition}\} := \text{among, between, by, of, with, to};

• Matched text:

PMID 16494873, SENT_ID 16494873_114

Upon precipitation of the SOX10 protein with anti-HA antibody, Western blot detection revealed expression of UBC9-V5 (25 kDa) in the sample (Fig. 1, line 6), indicating that \{UBC9 was co-immunoprecipitated with SOX10\}.

INTERACTOR_1: UBC9 resolved to UniprotID: UBC9_RAT
INTERACTOR_2: SOX10 resolved to UniProtID: SOX10_RAT
\{c-interact\} := [UBC9_RAT]_{interactor_1}, [SOX10_RAT]_{interactor_2}
Biological Language Processing
TECHNOLOGY
Unstructured Information Management Architecture

- a software architecture for developing and deploying *unstructured information management (UIM) applications*

- **UIM application**: a software system
  - analyse large volumes of unstructured information to
    - discover,
    - organize, and
    - deliver relevant knowledge to the end user

- software architecture which specifies
  - component interfaces, data representations, …

- Apache project: [http://uima.apache.org](http://uima.apache.org)
UIMA: Unstructured Information Management Architecture

Collections & Meta data → Collection Reader → CAS Initializer → Analysis Engine (adds to CAS)

... may be used by a Collection Reader to populate a CAS from a document. An example of a CAS Initializer is an HTML parser that de-tags an HTML document and also inserts paragraph annotations (determined from <P> tags in the original HTML) into the CAS.

... consume the enriched CAS that was produced by the sequence of Analysis Engines before it, and produce an application-specific data structure, such as a search engine index or database.

The UIMA CAS

• Data structure storing all of the information relating to a document and its meta-data
  – Source document text
  – Annotations produced by analysis engines

• Annotation characteristics
  – Stand-off: text annotations defined in terms of text spans
  – Semantics: defined by a type system
The role of the UIMA type system

• Defines the kinds of meta-data that can be stored in the CAS

• Supports description of the behavior of processing modules in terms of input/output annotations

• Defines the semantics of artifact annotations that can be represented: the structure implicit in the unstructured artifact
Type System semantics

• Fully specified semantic model
  – Type system reflects domain semantics
  – Specific types for each relevant domain category

• Generic meta-model
  – Type system defines the structure of the domain semantics; how to characterize the domain semantics in UIMA
  – Semantic distinctions remain external to UIMA
UIMA type system as domain model
Type System abstractness

- fully specified semantic model
- basic semantic distinctions in TS; more detailed external
- fully abstract semantic model
Strategy:
Separate Annotation from Referent

• Annotation corresponds to text span
• Annotation occurrence occurrence of a Referent
• Referent points to external entity
Biological Language Processing

ANNOTATION

REPRESENTATION
Text-mined Concepts and the Semantic Web

• Where possible, we utilize concepts defined in external ontologies or knowledge bases

• The annotations corresponding to these concepts can be straightforwardly represented in e.g. the Annotation Ontology

\[
\begin{align*}
\text{mouse Mapk7} & \quad \text{kiao:denotesResource} \quad \text{EG:23939} \\
& \quad \text{hasLocation} \quad \text{rdfs:label} \\
& \quad \text{t}_3 \quad \text{“M. musculus Mapk7”}
\end{align*}
\]
In this study, we examined whether gp41-induced IL-10 up-regulation is mediated by the previously described synergistic activation of cAMP and NF-kappaB pathways. [PMID 10089566]
Higher-order relations

degradation of Mdm2 leads to increased transcriptional activity of p53

increase( cause: {Statement 1}, increased: {Statement 2} )

cause: { degradation (MGI:Mdm2) }
increased: { transcription (MGI:Trp53) }

experimental method: Method (PSI-MI)
location: Tissue (Foundational Model of Anatomy)

Statement 1
Statement 2
Linguistic Structures

regulation of transcription of Mapk7

POS Tagging
regulation/NN of/IN transcription/NN of/IN Mapk7/NN

Parse
(ROOT
  (NP
    (NP (NN regulation))
    (PP (IN of)
      (NP (NN transcription)))
    (PP (IN of)
      (NP (NN Mapk7)))))

Typed dependencies
prep(regulation-1, of-2)
pobj(of-2, transcription-3)
prep(regulation-1, of-4)
pobj(of-4, Mapk7-5)
Annotation Model Requirements

• Capture associations between text and both
  – Single concepts
  – Arbitrary sets of assertions
regulation of transcription of mouse \textit{Mapk7}
translation of interferon gamma protein

GO_0006412: translation

PRO:0000017: interferon gamma protein
Provenance

- store meta-data about annotations
- track sources of constituent parts
  - enable reasoning over confidence in an extraction
  - expose the basis (evidence) for an extraction

Annotation Level:
(_:a2 kiao:basedOnAnnotation _:a1)

Element Level:
(_:a2 kiao:mentionsStatementElement _:e1)
(_:e1 kiao:isObjectOf _:s2; kiao:basedOnResourceOf _:a1)
regulation of translation of interferon gamma protein

(_:a3 rdf:type kiao:StatementSetAnnotation)
(_:a3 kiao:mentionsStatement _:s3)
(_:a3 kiao:mentionsStatement _:s4)
(_:s3 rdf:subject _:r1;
  rdf:predicate rdfs:subClassOf;
  rdf:object go:GO_0006417)
(_:s4 rdf:subject _:r1;
  rdf:predicate go:regulates;
  rdf:object _:v1)
(_:a3 kiao:basedOnAnAnnotation _:a2)

(_:a3 kiao:mentionsStatementElement _:e3)
(_:e3 kiao:isObjectOf _:s4;
  kiao:basedOnSubjectOf _:s1)
Acknowledgements

- Larry Hunter (Lab director)
- Kevin Livingston (KaBOB)
- Mike Bada (Ontologist)
- Kevin Cohen (NLP)
- (Software engineers)
  - Bill Baumgartner
  - Chris Roeder
  - Tom Christiansen
- Other Lab members:
  - Helen Johnson, Hannah Tipney, Yuriy Malenkiy, Lynne Fox
- NIH grants
  - R01 LM 010120-01
  - R01 LM 009254
  - R01 LM 008111
  - R01 GM 083649
  - G08 LM 009639
  - T15 LM 009451