

NCIBI Literature Mining

Behind the Scenes, Web-Based Access

Alex Ade

National Center for Integrative Biomedical Informatics

University of Michigan

30 July, 2009



Introduction

- NCIBI Biomedical Literature and Accessory Databases
- Natural Language Processing Pipeline
- Database Access
- XML Querying

NCIBI Biomedical Literature and Accessory Databases

Source Pubmed Database

- Available from the National Center for Biotechnology Information (NCBI)
- Provides access to citations and abstracts from biomedical literature from ~1950 to present in ~5000 journals
- Publishers participating in Pubmed electronically submit their citations to NCBI prior to or at the time of publication
- Records are indexed with NLM's MeSH
- Abstracts only, not full-text

NCIBI Pubmed Database

- We lease the NLM Pubmed data set, extract the data from the source XML files, and populate our local NCIBI Pubmed database
- Updated nightly (~12K per update, but never more than 30K)
- ~20M documents
- ~19M unique documents
- ~11M abstracts

Source PMCOA Database

- Pubmed Central is the NIH's free digital archive of biomedical and life science literature
- Some articles in PMC are "open access" which means that an article is still protected by copyright, but is distributed under a Creative Commons or similar license that generally allows more liberal use than a traditional copyrighted work
- Full-text documents

NCBI PMCOA Database

- We download the PMCOA data set, extract the data from the source XML files, and populate our local NCIBI PMCOA database
- Updated monthly (~2K documents per update)
- ~107K documents

Accessory Databases

- Entrez Gene - is NCBI's database for gene-specific information
- Journal - all Entrez journals
- Taxonomy - contains names of all organisms represented in the genetic database with at least one nucleotide or protein sequence
- UMLS - Metathesaurus and Semantic Network, very large vocabulary database that contains information about biomedical and health related concepts, their names, and the relationships among them (GO, FMA, ICD9, etc.)
- MeSH - NLM's controlled vocabulary for indexing Pubmed

MeSH

- Medical Subject Headings
- The NLM's controlled vocabulary thesaurus
- Sets of naming descriptors in an eleven-level hierarchical structure that permits searching at various levels of specificity
- ~25K descriptors, 83 qualifiers, ~186K entry terms (synonyms)
- e.g. Ascorbic Acid (descriptor), administration & dosage (qualifier), Vitamin C (entry term), etc.

Natural Language Processing Pipeline

NLP Tools and Pipeline

- We “value add” to the Pubmed and PMCOA databases by doing Natural Language Processing to the data and adding those results to the database
- Sentence segmentation (boundary detection) using MxTerminator
- Named entity tagging using NameTagger, SciMiner, and Genia
- Part of speech, stemming, lemmatizing using Porter-Stemmer and StanfordParser
- Parsing using StanfordParser
- Information extraction using ClairLib and GIN-IE

MxTerminator

- Java based sentence boundary detector
- Written by Adwait Radnaparkhi
- No source code available
- Performs with an accuracy of 93.5% on Medline abstracts with the default model
- The version I use has been trained for biomedical text and has an accuracy of about 96%
- ~92M Pubmed sentences, ~16M PMCOA sentences

Named Entity Tagging

- Looking for important biomedical words or phrases in a sentence
- NameTagger - uses a dictionary approach to tagging gene/protein names and MeSH terms
- SciMiner - uses a dictionary approach to tagging gene/protein and metabolite names
- Genia - statistical tagger trained on 2000 Medline abstracts for tagging protein, DNA, RNA, cell line, and cell type names
 - Hand annotated abstracts selected using *human*, *blood cell*, and *transcription factor* MeSH terms

Ambiguous Tags

- There are Gene symbols and synonyms that are also common words
- E.g. be, is, was, to, and, in, for, not, on, with, he, as, you, do, at, with, this, her, she, an, my, if...
- Mask system based on capitalization, punctuation, gene2pubmed, etc.

POS, Stemming, Lemmatizing

- Determine the part-of-speech and stem or root of a word
- Using Porter-Stemmer and the StanfordParser
- Stemming mainly involves suffix stripping rules
- Lemmatizing applies rules based on POS to determine the lemma or root of a word so is highly dependent on accurate POS
- E.g. “run” is the stem of “running” and “runs,” but “ran” is the stem of “ran”
- E.g. “run” is the lemma of “running,” “runs,” and “ran”

Parsing

- Using the StanfordParser
- Phrase Structure parse and Typed Dependency parse are different ways of representing the structure of a sentence
- Phrase Structure
 - Represents nesting of multi-word constituents (chunks)
- Typed Dependencies Collapsed
 - Represents dependencies between individual words and labels those dependencies with grammatical relations (collapsed prepositions and conjunctions)

Phrase Structure

- Colony-stimulating factor-1 (CSF-1) is essential for macrophage growth, differentiation and survival.

```
(ROOT
  (S
    (NP
      (NP (JJ Colony-stimulating) (NN factor-1))
      (PRN (-LRB- -LRB-))
      (NP (NN CSF-1))
      (-RRB- -RRB-)))
    (VP (VBZ is)
      (ADJP (JJ essential)
        (PP (IN for)
          (NP
            (ADJP (NN macrophage) (NN growth))
            (, ,) (NN differentiation)
            (CC and)
            (NN survival))))))
    (. .)))
```

S: Simple Declarative Clause

NP: Noun Phrase

JJ: Adjective

NN: Noun

PRN: Parenthetical

VP: Verb Phrase

VBZ: Verb, 3rd person singular present

ADJP: Adjective Phrase

PP: Prepositional Phrase

IN: Preposition

CC: Coordinating Conjunction

Typed Dependencies Collapsed

- Colony-stimulating factor-1 (CSF-1) is essential for macrophage growth, differentiation and survival.

```
amod(factor-1-2, Colony-stimulating-1)
nsubj(essential-7, factor-1-2)
appos(factor-1-2, CSF-1-4)
cop(essential-7, is-6)
prep_for(essential-7, macrophage-9)
dep(macrophage-9, growth-10)
conj_and(macrophage-9, differentiation-12)
conj_and(macrophage-9, survival-14)
```

```
amod: adjectival modifier
nsubj: nominal subject
appos: appositional modifier
cop: copula
prep_for: prepositional modifier
conj_and: conjunct
dep: dependent
```

GIN-IE

- Uses the sentences, tags, and parse data to extract and annotate protein-protein interactions
- Results are available as an RSS feed (recently published documents) and within MiMI (soon to be all documents)
- E.g. The partial results of GIN-IE on sentences tagged for BRCA1
 - We finally demonstrated by immunoprecipitation of ACCA in cells, that the whole BRCA1 protein interacts with ACCA when phosphorylated on Ser1263.
 - In contrast, BRCA1-p53 interaction is weak or other mechanisms operate.
 - We documented an in vivo association of the endogenous BRCA1 with PR isoforms A and B and a direct in vitro interaction between BRCA1 and PR, which was partially mapped.
 - BRCA1 colocalized with TRF2 in telomerase-positive cells and with a small subset of ALT-associated PML bodies (APBs) in ALT cells.

GIN-IE RSS Feed

- <http://gin.ncibi.org/rss/gin-ie/interactions.rss>
- RSS 2.0 format
- Contains only the interactions from recently published documents in Pubmed
- ~4700 interactions in the feed

GIN-IE Data in MiMI

Gene Details
Molecule Details for Gene Entry **BRCA1 (GeneId: 672)** - [show/hide](#)

breast cancer 1, early onset

BRCA1 (Homo sapiens)

- Gene Type: protein-coding
- Chromosome: 17
- Map Locus: [17q21](#)
- Locus Tag: null

Other Names...

- BRCA1
- BRCA1
- BRCC1
- IRIS
- PSCP
- RNF53

Descriptions...

- Authorized Gene Description:** breast cancer 1, early onset
- Other descriptions...**
 - BRCA1/BRCA2-containing complex, subunit 1
 - breast and ovarian cancer susceptibility protein 1

Gene Attributes

Cellular Components...

- [BRCA1-BARD1 complex](#)
- [cellular component](#)
- [gamma-tubulin ring complex](#)
- [intracellular](#)
- [nucleus](#)
- [ubiquitin ligase complex](#)

Biological Processes...

- [DNA damage response, signal transduction by p53 class mediator resulting in transcription of c21 class mediator](#)
- [DNA damage response, signal transduction resulting in induction of apoptosis](#)
- [DNA repair](#)
- [androgen receptor signaling pathway](#)
- [cell cycle](#)
- [cell cycle checkpoints](#)
- [chromosome segregation](#)
- [double-strand break repair via homologous recombination](#)
- [fatty acid biosynthetic process](#)
- [negative regulation of cell cycle](#)
- [negative regulation of centriole replication](#)
- [negative regulation of fatty acid biosynthetic process](#)
- [negative regulation of transcription](#)
- [positive regulation of DNA repair](#)
- [positive regulation of protein ubiquitination](#)
- [positive regulation of transcription, DNA-dependent](#)
- [postreplication repair](#)
- [protein ubiquitination](#)
- [regulation of apoptosis](#)
- [regulation of cell proliferation](#)
- [regulation of transcription from RNA polymerase II promoter](#)
- [regulation of transcription from RNA polymerase III promoter](#)
- [response to estrogen stimulus](#)

Molecular Functions...

- [DNA binding](#)
- [androgen receptor binding](#)
- [enzyme binding](#)
- [metal ion binding](#)
- [molecular function](#)
- [protein binding](#)
- [transcription coactivator activity](#)
- [tubulin binding](#)
- [ubiquitin-protein ligase activity](#)
- [zinc ion binding](#)

Protein Interactions (179 gene interactions found/39 NLP interactions found) - [show/hide](#)

Protein Interactions (179 gene interactions found/39 NLP interactions found) - [show/hide](#)

View BRCA1 With Other NCIBI Tools: [GeneZMeSH](#) [Cytoscape](#) [Netbrowser](#) [GIN](#) [MiSearch](#)

GIN-IE Data in MiMI

University of Michigan MiMI Gene Details

39 nlp interactions found, displaying page 1 of 2.
[First/Prev] 1, 2 [Next/Last]

Gene1	Gene2	Taxid	Interaction Type	Sentence	Pubmed Id	See Mined Text
BRCA1 protein	ACCA	9606	interacts	We finally demonstrated by immunoprecipitation of ACCA in cells, that the whole BRCA1 protein interacts with ACCA when phosphorylated on Ser1263.	16698035	view
BRCA1	p53	9606	interaction	In contrast, BRCA1-p53 interaction is weak or other mechanisms operate.	17161371	view
BRCA1	PR	9606	interaction	The BRCA1-PR interaction has functional consequences.	16109739	view
BRCA1	HIF-1alpha	9606	interaction	An interaction between BRCA1 and HIF-1alpha was found in human breast cancer cells.	16543242	view
BRCA1	FOXO1	9606	activated	In summary, we identified a FOXO1 binding site within the BRCA1-responsive element of the p27(Kip1) promoter and showed that FOXO1 activated the promoter alone and in conjunction with BRCA1.	16321276	view
BRCA1	WRN	9606	interaction	The interaction between WRN and BRCA1 increases in cells treated with DNA cross-linking agents.		
BRCA1	TRF2	9606	colocalized	BRCA1 colocalized with TRF2 in telomerase-positive cells and with a small subset of ALT-associated PML bodies (APB).		
BRCA1	PR	10090	interaction	The BRCA1-PR interaction has functional consequences.		
BRCA1	FANCD2	9606	interacts	As FANCD2 interacts with BRCA1, is expressed in proliferating normal breast cells, and FANCD2 knockout mice die, we investigated the expression of FANCD2 in sporadic and hereditary invasive breast cancer patients to evaluate its possible role in breast cancer.		
BRCA1	TRF2	9606	colocalized	BRCA1 colocalized with TRF2 in telomerase-positive cells and with a small subset of ALT-associated PML bodies (APB).		
BRCA1	BARD1	9606	interaction	Immunoblot analysis shows that inhibition of BRCC36 has no effect on the activation of ATM, expression of p21 and expression of BARD1 following IR exposure.		
BRCA1	TRRAP	9606	interaction	M. (1996) Nature 382, 678-679), caused the loss of physical interaction between BRCA1 and TRRAP and significant reduction of BRCA1 transcription function by hGCN5/TRRAP.		
BRCA1	PR	9606	interaction	We documented an in vivo association of the endogenous BRCA1 with PR isoforms A and B and a direct in vitro interaction between BRCA1 and PR which was partially mapped.		
BRCA1	COBRA1	9606	interacted	Endogenously expressed COBRA1 interacted with the nuclear protein BRCA1 in human breast cancer cells.		
BRCA1	BARD1	9606	interaction	Transfection of BRCA1 N-terminal peptides that disrupted the cellular BRCA1-BARD1 interaction caused a loss of BRCA1-mediated transcriptional activation and increased apoptosis in single cell assays, but did not alter localization or expression of endogenous BARD1.		
BRCA1	BARD1	9606	interacts	BRCA1 interacts with BARD1 to generate significant ubiquitin ligase activity which catalyzes nontraditional Lys-6-linked polyubiquitination of substrates.		

NLP Derived Interactions

39 nlp interactions found, displaying page 1 of 2.
[First/Prev] 1, 2 [Next/Last]

Gene1	Gene2	Taxid	Interaction Type	Sentence
BRCA1 protein	ACCA	9606	interacts	We finally demonstrated by immunoprecipitation of ACCA in cells, that the whole BRCA1 protein interacts with ACCA when phosphorylated on Ser1263.
BRCA1	p53	9606	interaction	In contrast, BRCA1-p53 interaction is weak or other mechanisms operate.
BRCA1	PR	9606	interaction	The BRCA1-PR interaction has functional consequences.

Database Access

NCIBI NLP Web Service

- NCIBI NLP web service provides programmatic access to the Pubmed and PMCOA databases
- Loosely modeled after NCBI's eUtils because many biomedical researchers are familiar with that service
- Build a URL based query using a base URL and a set of parameters
- Returns an XML result set
- <http://nlp.ncibi.org/about.html>

NCIBI-WS Future Work

- Query for SciMiner tags in Pubmed
- Query for Nametagger and SciMiner tags in PMCOA
- Add Metadata to the Article (e.g. authors, title)
- Support for multiple, comma separated identifiers in a single query (e.g. pmid=1234,5678,9101)
- Add support for querying GIN-IE interaction data

XML Result (PMID)

<http://nlp.ncibi.org/fetch.php?pmid=17523140>

```
<?xml version="1.0"?>
<NCIBI>
  <BioNLP>
    <Request type="fetch">
      <ParameterSet>
        <PMID>17523140</PMID>
        <Limit>1000</Limit>
      </ParameterSet>
    </Request>
    <Response>
      <ResultSet>
        <Result>
          <Article pmid="17523140">
            <Section type="abstract">
              <Paragraph>
                <Sentence>Heat stress causes severe constraints on numerous physiological...</Sentence>
                <Sentence>In this study, we performed proteomic profiling of a nuclear...</Sentence>
                <Sentence>We found 10 protein spots whose expression had changed after heat...</Sentence>
                <Sentence>Seven of those protein spots, periodic tryptophan protein 1 homolog...</Sentence>
                <Sentence>We focused on the downregulation of two splicing factors, which might...</Sentence>
              </Paragraph>
            </Section>
          </Article>
        </Result>
      </ResultSet>
    </Response>
  </BioNLP>
</NCIBI>
```

XML Result (Tagged PMID)

<http://nlp.ncibi.org/fetch.php?pmid=17523140&tagger=nametagger&type=gene>

```
<?xml version="1.0"?>
<NCIBI>
  <BioNLP>
    <Request type="fetch"><ParameterSet><PMID>17523140</PMID><Limit>1000</Limit></ParameterSet></Request>
    <Response><ResultSet><Result>
      <Article pmid="17523140">
        <Section type="abstract">
          <Paragraph>
            <Sentence>Heat stress causes severe constraints on numerous physiological...</Sentence>
            <Sentence>In this study, we performed proteomic profiling of a nuclear...</Sentence>
            <Sentence>We found 10 protein spots whose expression had changed after heat...</Sentence>
            <Sentence>Seven of those protein spots, <Gene type="entrez" id="11137">periodic tryptophan protein 1</Gene> homolog (<Gene type="entrez" id="11137">PWP1</Gene>), <Gene type="entrez" id="3837">importin beta-1 subunit</Gene>, sumoylated protein, <Gene type="entrez" id="10946">splicing factor 3a subunit 3</Gene> (<Gene type="entrez" id="10946">SF3a3</Gene>), <Gene type="entrez" id="23435">TAR DNA-binding protein 43</Gene>, <Gene type="entrez" id="26854,6066,26853,26855">U2</Gene> small nuclear ribonucleoprotein auxiliary factor 35 kDa subunit (<Gene type="entrez" id="7307">U2AF35</Gene>) and small ubiquitin-related modifier-1 (<Gene type="entrez" id="7341">SUMO-1</Gene>) were downregulated, while three other protein spots, Protein SET, 40S <Gene type="entrez" id="3921">ribosomal protein SA</Gene> and <Gene type="entrez" id="6175">60S acidic ribosomal protein P0</Gene> were upregulated by the heat stress.</Sentence>
            <Sentence>We focused on the downregulation of two splicing factors, which might...</Sentence>
          </Paragraph>
        </Section>
      </Article>
    </Result></ResultSet></Response>
  </BioNLP>
</NCIBI>
```

XML Result (Gene ID)

<http://nlp.ncibi.org/fetch.php?tagger=nametagger&type=gene&id=11137>

```
<?xml version="1.0"?>
<NCIBI>
  <BioNLP>
    <Request type="fetch">
      <ParameterSet><Tagger>nametagger</Tagger><Type>gene</Type><ID>11137</ID><Limit>1000</Limit></ParameterSet>
    </Request>
    <Response>
      <ResultSet>
        <Result>
          <Article pmid="7828893">
            <Section type="abstract">
              <Paragraph>
                <Sentence>We have cloned and expressed in vaccinia virus a cDNA encoding an...</Sentence>
                <Sentence>Database searching indicated that <Gene type="entrez" id="11137">IEF SSP...</Sentence>
              </Paragraph>
            </Section>
          </Article>
        </Result>
        <Result>
          <Article pmid="11850830">
            <Section type="abstract">
              <Paragraph>
                <Sentence>The transcript encoding endonuclein, the human homolog of yeast <Gene...</Sentence>
              </Paragraph>
            </Section>
          </Article>
        </Result>
        ...
      </ResultSet>
    </Response>
  </BioNLP>
</NCIBI>
```

XML Querying

XPath

- Defined by the World Wide Web consortium
- XPath is a query language for XML
- XPath is used to navigate through the nodes and attributes in an XML document
- Implementations available for the major bioinformatics programming languages (e.g. Perl, Python, Java)

XPath Example (NCIBI)

```
URL ncibiws = new URL("http://nlp.ncibi.org/fetch.php?pmid=17523140&tagger=nametagger&type=gene");
URLConnection connection = ncibiws.openConnection();

InputStream inputStream = connection.getInputStream();

DocumentBuilderFactory factory = DocumentBuilderFactory.newInstance();
Document document = factory.newDocumentBuilder().parse(inputStream);

inputStream.close();

XPath xpath = XPathFactory.newInstance().newXPath();
String expression = "//Sentence";

NodeList nodes = (NodeList)xpath.evaluate(expression, document, XPathConstants.NODESET);

for (int i = 0; i < nodes.getLength(); i++) {
    System.out.println(nodes.item(i).getTextContent());
}
```

XPath Example (NCIBI + eUtils)

```
URL entrezws =
    new URL("http://eutils.ncbi.nlm.nih.gov/entrez/eutils/esearch.fcgi?db=pubmed&term=ade%20a[au]");

// will return a nodelist of PMIDs where I'm an author
String expression = "//IdList/Id";

NodeList pmidNodes = (NodeList)xpath.evaluate(expression, document, XPathConstants.NODESET);

for (int i = 0; i < pmidNodes.getLength(); i++) {
    String PMID = pmidNodes.item(i).getTextContent();

    URL ncibiws = new URL("http://nlp.ncibi.org/fetch.php?pmid=" + PMID);

    // will return a nodelist of Sentences that have a Gene tag
    expression = "//Gene/parent:node()";

    NodeList sentenceNodes = (NodeList)xpath.evaluate(expression, document, XPathConstants.NODESET);

    for (int k = 0; k < sentenceNodes.getLength(); k++) {
        System.out.println(sentenceNodes.item(i).getTextContent());
    }
}
```


More XPath Expressions

```
// pubmed
```

```
// parameters pmid, tagger, type
```

```
String expression = “//Sentence”;
```

```
String expression = “//Gene”;
```

```
String expression = “//Gene/parent::node()”;
```

```
String expression = “//Gene[@id=‘11137’]/parent::node()”;
```

```
// parameters tagger, type, id
```

```
String expression = “//Article/@pmid”;
```

```
// pmcoa (soon)
```

```
// parameters pmcid, tagger, type
```

```
String expression = “//Section[@type=‘results’]//Gene/parent::node()”;
```

Thank You!

asade@umich.edu
nlp-help@umich.edu