Key Concepts in Text Mining and its Usage in Service Provision

Jee-Hyub Kim and Şenay Kafkas
Literature Services Team
EMBL-EBI

23 March 2016
Contents

● Session 1
  ○ Introduction to text mining
  ○ Methods, Available tools, Document and format types, Gold standard data, Performance evaluation

● Session 2
  ○ Europe PubMed Central: Finding articles
  ○ Building your own pipeline
Contents

● Session 3
  ○ An industrial project as sample: Europe PMC CTTV literature component
    ■ Project overview
    ■ Methods
    ■ Results
    ■ Future Work
Session 1
Why do we need text mining?
Why do we need text mining?

● Everyday we use Google.
● Google helps us to find right information (articles) or leads us to a right place (website) for further search.
  ○ Amazon: a product-specific search engine
  ○ PubMed: scientific articles
● Sometimes, Google is not good enough for what you’re searching for.
● For example,
Questions Google can’t answer

● Could you give me a summary instead of me reading a large number of articles?
● Finding articles where genus Cancer used in Methods section.
● Finding data linked to articles.
● Can I find articles where gene-disease relations are mentioned?
● Can you find all the articles where cell lines used in 1000 Genomes mentioned in Methods?
An example answered by text mining

- Finding articles where genus Cancer used in Methods section.
- On Europe PMC
  - “Cancer” 4,042,737
  - (METHODS:"Cancer") AND (ORGANISM:"Cancer")
    AND (IN_EPMC:Y OR HAS_BOOK:Y) -> 90 articles
- Text mining can help you find those 90 rare articles from 4,042,737.
There are some text mining tools already available for you.

- In order to find a right tool, you need to define your problem.
- Having a bit of text-mining knowledge will help you to define your problems.
- Later, we will show you a list of tools for text mining tasks.
What is text mining?
What is text mining?

• “Text Mining is the discovery by computer of new, previously unknown information, by automatically extracting information from different written resources.” (Marti Hearst)
• More practical definition: helps you to find right documents and extract information automatically. If possible, combine extracted information leading to new findings.
  ■ Information Retrieval
  ■ Information Extraction
  ■ Data Mining
We are all familiar with information retrieval and in fact we google everyday.

Also in life science:
- Everyday, a large number of new documents are published and indexed.
- For biologists, PubMed is a popular search tool.
Information Retrieval: Search Algorithms

- **Boolean model**: AND, OR, NOT
  - e.g., PubMed

- **Vector space model**: a bag of words
  - e.g., Europe PMC

- **PageRank**
  - e.g., Google

- **Faceted search (document categorization)**
  - e.g., amazon.com
A text analysis method that transforms unstructured text to structured outputs (e.g., templates)

Templates define what to extract
- Similar to a table.

Information extraction tasks
- Named Entity Recognition
- Relation Extraction
- Event Extraction (a hard problem)
Named Entity Recognition
e.g., Francis Crick is a researcher.
Heme oxygenase-1: a novel therapeutic target for *gastrointestinal* diseases.

1. **Abstract**

Heme oxygenase-1 (HO-1) is the rate-limiting enzyme in the **catabolism** of heme, followed by production of biliverdin, free iron and carbon monoxide (CO). HO-1 is a stress-responsive protein induced by various oxidative agents. Recent studies demonstrate that the expression of HO-1 in response to different inflammatory mediators may contribute to the resolution of inflammation and has protective effects in several organs against oxidative injury. Although the mechanism underlying the anti-inflammatory actions of HO-1 remains poorly defined, both CO and biliverdin/bilirubin have been implicated in this response. In the gastrointestinal tract, HO-1 is shown to be transcriptionally induced in response to oxidative stress, preconditioning and acute inflammation. Recent studies suggest that the induction of HO-1 expression plays a critical protective role in intestinal damage models induced by ischemia-reperfusion, Indomethacin, lipopolysaccharide-associated sepsis, trinitrobenzene sulfonic acid, and dextran sulfate sodium, indicating
Relation Extraction

What proteins interact with BRCA2?

215 binary interactions found for search term **BRCA2**

Network visualisation
What’s Coming Next?

- We introduced:
  - What is text mining?
  - Text-mining tasks
- In the next slides, we will introduce some background knowledge to help you find right tools and develop your own tools
- Challenges in biomedical domain: A long tail
- Various document types and formats
- It’s also good to know text-mining methods to build your own tool.
Methods, Tools and Services, Document and format types, Gold Standards and Performance Evaluation in Information Extraction tasks
TM Methods

● Dictionary-based method
  ○ NER
  ○ Context-independent
  ○ Easily linked to DB/Ontology entries
  ○ Dictionaries are generally complementary, so it would be good to merge them before using!

● Machine Learning-based method
  ○ Many applications (NLP tools, NER, Relationship Extraction, article classification)
  ○ Context-dependent
  ○ Gold standards (annotated corpora) are needed
TM Methods

- Rule-based method
  - Many applications (e.g. NER, Relationship Extraction, article classification)
  - Simple $\rightarrow$ complex rules
    - Co-occurrences of entities
    - Regular Expressions
    - Heuristic data
    - NLP tools

- Combining different methods
  - E.g. dictionary-based + ML-based
## Some Tools and Services are available

<table>
<thead>
<tr>
<th>Tool Type</th>
<th>Tool Name</th>
<th>Licence Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tokenizer</td>
<td>GENIA tagger</td>
<td>None</td>
</tr>
<tr>
<td>Sentence splitter</td>
<td>LingPipe</td>
<td>Commercial</td>
</tr>
<tr>
<td>Named entity tagger</td>
<td>Banner</td>
<td>Public</td>
</tr>
<tr>
<td>Section Tagger</td>
<td>Europe PMC section tagger</td>
<td>Public</td>
</tr>
<tr>
<td>Protein-protein interaction extractor</td>
<td>iHop</td>
<td>Public</td>
</tr>
<tr>
<td>Platform for tool development</td>
<td>UIMA</td>
<td>Apache License v2.0</td>
</tr>
<tr>
<td>Package for tool development</td>
<td>WEKA</td>
<td>Public</td>
</tr>
</tbody>
</table>
# How to choose/develop the right tool?

<table>
<thead>
<tr>
<th>Choose a tool</th>
<th>Develop a tool</th>
</tr>
</thead>
<tbody>
<tr>
<td>Good understanding of the problem definition is the key</td>
<td>Is there a tool that is easy to adapt it to my problem?</td>
</tr>
<tr>
<td>Does its performance meet the requirements?</td>
<td>Which method should I use?</td>
</tr>
<tr>
<td>Is it easy to use?</td>
<td></td>
</tr>
<tr>
<td>Can this tool operate on the document collection that I want to analyse?</td>
<td></td>
</tr>
<tr>
<td>How long this service/tool would be available for me and is it up-to-dated</td>
<td></td>
</tr>
<tr>
<td>regularly?</td>
<td></td>
</tr>
</tbody>
</table>
Different document collections, different formats...

- **Document collections:**
  - Abstracts
  - Full text articles
  - Electronic Health Records
  - Patents
  - Social Media
  - ...

- **Document formats**
  - Text
  - PDF
  - XML
  - ...
  - ...
Performance Evaluation - Metrics

- **Precision**: exactness – what % of samples that the system identified as positive are actually positive.
- **Recall**: completeness – what % of positive samples did the system identify as positive?
- **F-Score**: Harmonic mean of Precision and Recall.

\[
P = \frac{TP}{TP + FP} \]

\[
R = \frac{TP}{TP + FN} \]

\[
F-score = \frac{2 \times P \times R}{P + R} \]

- **Automatic**
  - Availability of gold standards
    - expert curated document collections
    - Expensive in terms of time (producing guidelines + curation)

- **Manual**
  - Random selection of a set of results and manual evaluation of the selected samples by the help of domain expert(s).
What are the parameters affecting system performance?

- Implementation methodologies
  - Dictionary based: Is my dictionary up-to-dated? Does it have a good coverage? Does it contain many generic, ambiguous names?
  - Rule based: Are my rules up-to-dated and complete?
  - ML based: Did I choose the correct algorithm, set of features and parameters to train my system?
What are the parameters affecting system performance?

- Gold standard corpus (If you use any)
  - Quality of the annotations?
    - Does it contain generic annotations?
    - Are the annotations in the corpus up-to-dated?
  - Does it represent a narrow/board domain?
  - Does it reflect the distribution in the real text collection?
  - Size of the Corpus
  - Distribution of the sample classes in the corpus
Take Home Messages

● Each real world problem has its own requirements
● Good understanding of the problem requirements is the key for choosing or developing a TM tool
● TM is challenging
  ○ Complexity of natural language
  ○ Ambiguous named entities (with common English words, within and between species)
  ○ Big data
  ○ ...
● TM can help you only to alleviate your problems
  ○ Computers cannot mimic human perception and TM deals with unstructured data which requires human perception
Session 2
Europe PMC: Searching for Articles
Europe PMC Overview

- 30+ million documents of which 3+ million full text and >25 million PubMed abstracts
- 1M Open Access
- Enrichments: ORCIDs, citations, named entities, data links, section level search
- Website, web services, FTP
- 29 funders including EMBL
- Most widely used service of EBI: Over 1M IPs per month
Searching Articles by Relevance, Date, Citation
1. **In vitro and in vivo characterization of new swine-origin H1N1 influenza viruses.**
   - Cited: 506 times

2. **The Universal Protein Resource (UniProt) in 2010.**
   - Authors: UniProt Consortium
   - Cited: 494 times

3. **Structural and functional bases for broad-spectrum neutralization of avian and human influenza A viruses.**
   - Cited: 431 times

4. **Experimental adaptation of an influenza H5 HA confers respiratory droplet transmission to a reassortant H5 HA/H1N1 virus in ferrets.**
   - Cited: 401 times

5. **Global spread of Carbapenemase-producing Enterobacteriaceae.**
   - Authors: Nordmann P, Naas T, Poirel L
   - Cited: 359 times
Structural and functional bases for broad-spectrum neutralization of avian and human influenza A viruses.

(PMD:19234466 PMCID:PMC2692245)


Department of Cancer Immunology & AIDS, Dana Farber Cancer Institute, Harvard Medical School, 44 Binney Street JFB 826, Boston, Massachusetts 02115, USA. jianhua_sui@dfci.harvard.edu


Type: Journal Article, Research Support, N.I.H., Extramural

DOI: 10.1038/nsmb.1566

Abstract

Influenza virus remains a serious health threat, owing to its ability to evade immune surveillance through rapid genetic drift and reassortment. Here we used a human non-immune antibody phage-display library and the H5 hemagglutinin ectodomain to select ten neutralizing antibodies (nAbs) that were effective against all group 1 influenza viruses tested, including H5N1 "bird flu" and the H1N1 "Spanish flu." The crystal structure of one such nAb bound to H5 shows that it blocks infection by inserting its heavy chain into a conserved pocket in the stem region, thus preventing membrane fusion. Nine of the nAbs employ the germline gene VH1-69, and all seem to use the same neutralizing mechanism. Our data further suggest that this region is recalcitrant to neutralization escape and that nAb-based immunotherapy is a promising strategy for broad-spectrum protection against seasonal and pandemic influenza viruses.

Funding

NCRR NIH HHS (P41 RR-0181)

NIAD NIH HHS (P01 AI055789, P01 AI055789-040001, U01 AI07518-01, P01 AI055789)

Read Article at publisher's site
Abstract

Influenza virus remains a serious health threat, owing to its ability to evade immune surveillance through rapid genetic drift and reassortment. Here we used a human non-immune antibody phage-display library and the H5 hemagglutinin ectodomain to select ten neutralizing antibodies (nAbs) that were effective against all group 1 influenza viruses tested, including H5N1 'bird flu' and the H1N1 'Spanish flu'. The crystal structure of one such nAb bound to H5 shows that it blocks infection by inserting its heavy chain into a conserved pocket in the stem region, thus preventing membrane fusion. Nine of the nAbs employ the germline gene VH1-69, and all seem to use the same neutralizing mechanism. Our data further suggest that this region is recalcitrant to neutralization escape and that nAb-based immunotherapy is a promising strategy for broad-spectrum protection against seasonal and pandemic influenza viruses.

Funding

NCRR NIH HHS [P41 RR-01081]
NIAID NIH HHS [P01 AI055789, P01 AI055789-040001, U01-AI07518-01, P01-AI055789]

Read Article at publisher's site
Named Entity Summary Table

**Genes & Proteins**

<table>
<thead>
<tr>
<th>Gene/Protein</th>
<th>UniProt ID</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hemagglutinin</td>
<td>UniProt:Q7X8G2</td>
</tr>
<tr>
<td>Hemagglutinin</td>
<td>UniProt:Q6J8B6</td>
</tr>
</tbody>
</table>

**Protein Families**

<table>
<thead>
<tr>
<th>Family</th>
<th>InterPro ID</th>
</tr>
</thead>
<tbody>
<tr>
<td>Capsid hemagglutinin</td>
<td>InterPro:IPR003956</td>
</tr>
<tr>
<td>Hemagglutin HA1 beta-hem. dom</td>
<td>InterPro:IPR012672</td>
</tr>
<tr>
<td>Hemagglutin influenza A</td>
<td>InterPro:IPR001919</td>
</tr>
<tr>
<td>Hemagglutin influenza A/B</td>
<td>InterPro:IPR001364</td>
</tr>
</tbody>
</table>

**Protein Structures**

<table>
<thead>
<tr>
<th>Structure</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 PDBe structure</td>
</tr>
</tbody>
</table>

*Structural and functional bases for broad-spectrum neutralization of avian and human influenza A viruses.*

(PMID:19234466 PMCID:PMC2692245)
Semantic Search Powered by NER

Search worldwide, life-sciences literature

(METHODS:“Cancer”) AND (ORGANISM:“Cancer”) AND OPEN_ACCESS:Y

E.g. “breast cancer” HER2 Smith J

Results

Sort by: Relevance | Date | Times Cited

Results 1 - 25 of 43

☐ A Vulnerability Assessment of Fish and Invertebrates to Climate Change on the Northeast U.S. Continental Shelf.
(PMID:26839567 PMCID:PMC4739546) Free full text article
Hare JA, Morrison WE, Nelson MW, Stachura MM, Teeters EJ, Griffis RB, Alexander MA, Scott JD, Alade L, Bell RJ, Chute AS, Curti KL, Curtis TH, Kirchels D, Kock JF, Lucey SM, McCandless CT, Milke LM, Richardson DE, Robillard E, ...
PLoS One [2016, 11(2)e0146756]
Cited: 0 times

☐ Projected Scenarios for Coastal First Nations’ Fisheries Catch Potential under Climate Change: Management Challenges and Opportunities.
(PMID:26761439 PMCID:PMC4711888) Free full text article
Weatherdon LV, Ota Y, Jones MC, Close DA, Cheung WW
PLoS One [2016, 11(1)e0145285]
Cited: 0 times

☐ Construction and analysis of IncRNA-IncRNA synergistic networks to reveal clinically relevant IncRNAs in cancer.
(PMID:26305574 PMCID:PMC4694810) Free full text article

Popular content sets
All results (90)
Full Text articles only (90)
Open Access articles only (43)
All reviews (4)
3. Structural comparison

A comparison of all the structures of prothrombin and fragments, both zymogen and enzyme, show several important structural rearrangements that the protein undergoes during activation [37], [38], [39], [40], [41], [42] (Fig. 3). Early structures of the human PPACK-thrombin/F2 complex (PDBID:2HPQ) and bovine PPACK-meizothrombin(desF1) (PDBID:1A0H) show similar binding interfaces for F2 on the catalytic domain, which we have termed ‘mode 1’ (red in the web version) in Fig. 4. The subsequent structure of human PPACK-meizothrombin(desF1) (PDBID:3E6P) confirmed this binding interface [43]. In all of these structures the catalytic domain has been activated by cleavage at Arg320 and the active site of the enzyme was covalently bound to the zymogen inhibitor (D-phenylalanyl-prolyl-arginyln chloromethylketone). An extensive contact surface between the catalytic domain is observed in these structures, burying a total of ~1600 Å² (Fig. 6). Several salt-bridges, hydrogen bonds, and hydrophobic contacts (Fig. 6). The interactions include: Arg223 and Arg225, Arg409 (93 in chymotrypsin numbering) and Arg416 with Arg490(165); Glu227 with Arg500(175); and Asp239 with Arg413(97). Other basic residues on the C-terminus, Arg565(233), Lys568(236) and Lys572(240), do not form salt-bridges, although they have been shown to be critical for binding other exosite II ligands. In terms of number and types of contacts, the most important residue in exosite II is Arg409(93). It participates in multiple salt-bridges, water-mediated hydrogen bonds, pi-stacking and hydrophobic interactions, and accounts for up to 18% of the buried surface area.
Mining Accession Numbers

Data Citations
Identified 7 unique Data Citations in the full text

pdb 3K5F (2)
pdb 1A0H (2)
pdb 2HPQ (1)
pdb 3E6P (1)

PDB STRUCTURE 3E6P

pdb 3NKP (1)
pdb 4H2H (1)
pdb 4NZQ (1)
Articles with data deposited in Dryad
Search worldwide, life-sciences literature

(TITLE:"biodiversity") AND (LABS_PUBS:"1012")

E.g. "breast cancer" HER2 Smith

Results

Sort by: Relevance | Date | Times Cited

Results 1 - 25 of 68

Marine Biodiversity in Juan Fernández and Desventuradas Islands, Chile: Global Endemism Hotspots.
(PMID:26734732 PMCID:PMC4703205) Free full text article
Friedland AM, Ballesteros E, Caselle JE, Gaymer CF, Palma AT, Pettit I, Varas E, Muñoz Wilson A, Sala E
PloS One [2016, 11(1)p0145059]
Cited: 0 times

Assessing vertebrate biodiversity in a kelp forest ecosystem using environmental DNA.
(PMID:26586544 PMCID:PMC4737306) Free full text article
Cited: 0 times

Analysing biodiversity and conservation knowledge products to support regional environmental assessments.
(PMID:28881749 PMCID:PMC4755129) Free full text article
Sci Data [2016, 3:160007]
Cited: 0 times

Popular content sets

Full Text articles only (31)
Open Access articles only (24)
All reviews (1)
RESTful Web Services

Articles RESTful API

RESTful Web Service

1. Quick Start
2. Web Service Overview
3. Sorting Results
4. Request Methods
5. Release Notes

Quick Start

For a quick start, click on the following link (opens in new window):

http://www.ebi.ac.uk/europepmc/webservices/rest/search?query=p53

Come back to this page for further guidance and help.

Web Service Overview

Our RESTful Web Service gives you access to all of the publications and related information we hold in the Europe PMC database, which includes the following:

- Over 30.4 million publications from various sources, including PubMed, AGRICOLA, and the European Patent Office.
- We provide the full text article free for more than 3.3 million of these publications, of which more than one million are open access.
- Database cross-references to a number of databases, including UniProt, the European Nucleotide Archive (ENA; formerly EMBL), the Protein Data Bank in Europe (PDBe), and more.
- Reference lists for more than 5 million publications.
- Citation counts and a citation network.
- Terms that have been text-mined from full text articles, including accession numbers, chemicals, diseases, genes and proteins, Gene Ontology terms, and organisms.
Read, Annotate, and Share
Thanks to text mining, relevant articles are found!

- Once relevant articles are found, we start reading.
- Often we highlight and annotate articles.
- Wouldn’t it be nice if we can share those annotations to discuss with your colleagues?
  - There are annotation tools for this.
- Your annotation also can help with improving text-mining algorithms, which helps with improving search results.
**Annotator**

**The copy-editor of the web**

Annotator is an open-source JavaScript library to easily add annotation functionality to any webpage. Annotations can have comments, tags, links, users, and more. Annotator is designed for easy extensibility so it's a cinch to add a new feature or behaviour. Annotator also fosters an active developer community with contributors from four continents, building 3rd party plugins allowing the annotation of PDFs, EPUBs, videos, images, sound, and more.

Download: 1.2.x stable  2.x pre-release  GitHub

---

**Using Annotator**

Simply select some text on this page and make a comment!

Adding annotation to any website is easy with Annotator. First, download the Annotator library, include it in your HTML:

Then, add the following javascript to initialize the annotator:

```javascript
var ann = Annotator(document.body);
```

**Plugins**

Annotator has a simple but powerful plugin architecture. Plug-ins can also be included for adding functionality such as user permissions, tags, filtering and formatting.

The Annotator community has developed more than twenty plugins, such as:

- Annotorous - image annotation
- Touch Plugin - support for touch devices (iOS, Android, ...
Heme oxygenase-1: a novel therapeutic target for gastrointestinal diseases

Yuji Naito,† Tomohisa Takagi, Kazuhiro Uchiyama, and Toshikazu Yoshikawa

Author information ► Article notes ► Copyright and License information ►

This article has been cited by other articles in PMC.

Abstract

Heme oxygenase-1 (HO-1) is the rate-limiting enzyme in the catabolism of heme, followed by production of biliverdin, free iron and carbon monoxide (CO). HO-1 is a stress-responsive protein induced by various oxidative agents. Recent studies demonstrate that the expression of HO-1 in response to different inflammatory mediators may contribute to the resolution of inflammation and has protective effects in several organs against oxidative injury. Although the mechanism underlying the anti-inflammatory actions of HO-1 remains poorly defined, both CO and biliverdin/bilirubin have been implicated in this response. In the gastrointestinal tract, HO-1 is shown to be transcriptionally induced in response to oxidative stress, preconditioning and acute inflammation. Recent studies suggest that the induction of HO-1 expression plays a critical protective role in intestinal damage models induced by ischemia-reperfusion, indomethacin, lipopolysaccharide-associated sepsis, trinitrobenzene sulfonic acid, and dextran sulfate sodium, indicating that activation of HO-1 may act as an endogenous defensive mechanism to reduce inflammation and tissue injury in the gastrointestinal tract. In addition, CO derived from HO-1 is shown to be involved in the regulation in gastro-intestinal motility. These in vitro and in vivo data suggest that HO-1 may be a novel therapeutic target in patients with gastro-intestinal diseases.
Building your own pipeline with a dictionary-based tagger
Example: Europe PMC Text-Mining Pipeline

- Full-text articles
- Section tagger
- Sentence splitter
- Dictionary-based taggers
- Annotated full-text articles

Controlled vocabularies, ontologies
<table>
<thead>
<tr>
<th>Use case</th>
<th>Semantic type</th>
<th>Text-mining task</th>
<th>Document type</th>
<th>Section</th>
</tr>
</thead>
<tbody>
<tr>
<td>CTTV</td>
<td>gene, disease</td>
<td>RE</td>
<td>abstract, article</td>
<td></td>
</tr>
<tr>
<td>KEW Garden</td>
<td>species (multilingual)</td>
<td>NER</td>
<td>article</td>
<td></td>
</tr>
<tr>
<td>Impact Assessment</td>
<td>accession number</td>
<td>NER</td>
<td>patent</td>
<td>Title, Claim, Description, Abstract, Figure, Table</td>
</tr>
<tr>
<td>ERC</td>
<td>grant identifier</td>
<td>NER</td>
<td>article</td>
<td>Acknowledgements</td>
</tr>
<tr>
<td>ELIXIR-EXCELERATE</td>
<td>resource name</td>
<td>NER</td>
<td>article</td>
<td></td>
</tr>
<tr>
<td>1000 Genomes</td>
<td>cell line</td>
<td>NER</td>
<td>article</td>
<td></td>
</tr>
<tr>
<td>Wikipedia</td>
<td>accession number</td>
<td>NER</td>
<td>wikipage</td>
<td></td>
</tr>
<tr>
<td>ChEMBL</td>
<td>resource name</td>
<td>NER</td>
<td>article</td>
<td></td>
</tr>
</tbody>
</table>
Dictionary-Based Taggers

- A key building block in a text-mining pipeline.
- We choose a dictionary-based approach for NER:
  - Simple and fast
  - Readable
  - Interactive
- A dictionary: a mapping table from terms to IDs.

<table>
<thead>
<tr>
<th>Term</th>
<th>ID</th>
</tr>
</thead>
<tbody>
<tr>
<td>inflammatory bowel disease</td>
<td>EFO_0003767</td>
</tr>
<tr>
<td>Inflammatory Bowel Diseases</td>
<td>EFO_0003767</td>
</tr>
<tr>
<td>IBD</td>
<td>EFO_0003767</td>
</tr>
</tbody>
</table>

- Two main sources: ontologies and spreadsheets
Building your own dictionary-based tagger

- **Monq.jfa**
  - Java class library for fast and flexible text filtering with regular expressions
  - Download from [www.pifpafpuf.de/Monq.jfa/](http://www.pifpafpuf.de/Monq.jfa/)

- **Building a dictionary**
  - Choose columns for terms and IDs from your table.
  - Create a mwt file.
    - a xml file for mapping from terms to IDs.
Building EFO Disease Dictionary

Experimental Factor Ontology

http://www.ebi.ac.uk/efo

The Experimental Factor Ontology (EFO) provides a systematic description of many experimental variables available in EBI databases, and for external projects such as the NHGRI GWAS catalogue. It combines parts of several biological ontologies, such as anatomy, disease and chemical compounds. The scope of EFO is to support the annotation, analysis and visualization of data handled by many groups at the EBI and as the core ontology for the Centre for Therapeutic Validation (CTTV)

Ontology info

Ontology Id: efo
Number of terms: 17892
Version: 2.69
Last loaded: Thu Feb 18 01:35:11 GMT 2016
Contact E-mail: efo-users@lists.sourceforge.net
Choose Subtrees and Properties

**disease**

- ID: EFO_0000249
  - Term: alzheimer disease
  - Semantic type: disease
  - Status: synonym

- ID: EFO_0003767
  - Term: inflammatory bowel disease
  - Semantic type: disease
  - Status: label

- ID: EFO_0003767
  - Term: IBD
  - Semantic type: disease
  - Status: synonym

**phenotype**

- The observable form taken by some character (or group)

**alternative_term**

- An alternative label for a given entity such as a commonly used abbreviation or synonym.
OLS API http://www.ebi.ac.uk/ols/beta/docs/api

Search

Search terms

The search API is independent of the REST API and supports free text search over the ontologies. The default search is across all textual fields in the ontology, but results are ranked towards hits in labels, then synonyms, then definitions, then annotations.

GET /api/search?q={query}

Search parameters

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>query</td>
<td>The terms to search. By default the search is performed over term labels, synonyms, descriptions, identifiers and annotation properties.</td>
</tr>
</tbody>
</table>

You can override the fields that are searched by supplying a queryFields argument. For example, to query on labels and synonyms use

GET /api/search?q={query}&queryFields={label,synonym}

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ontology</td>
<td>Restrict a search to a set of ontologies e.g. ontology=uberon,ma</td>
</tr>
<tr>
<td>type</td>
<td>Restrict a search to an entity type, one of {term,property,individual}</td>
</tr>
<tr>
<td>slim</td>
<td>Restrict a search to a particular set of slims by name</td>
</tr>
<tr>
<td>fieldsList</td>
<td>Specify the fields to return, the defaults are {iri,label,short_form,obo_id,ontology_name,ontology_prefix,description,type}</td>
</tr>
</tbody>
</table>
Building a dictionary

<table>
<thead>
<tr>
<th>ID</th>
<th>Term</th>
<th>Semantic type</th>
<th>Status</th>
</tr>
</thead>
<tbody>
<tr>
<td>EFO_0000249</td>
<td>alzheimer disease</td>
<td>disease</td>
<td>synonym</td>
</tr>
<tr>
<td>EFO_0003767</td>
<td>inflammatory bowel disease</td>
<td>disease</td>
<td>label</td>
</tr>
<tr>
<td>EFO_0003767</td>
<td>IBD</td>
<td>disease</td>
<td>synonym</td>
</tr>
</tbody>
</table>

```xml
<mwt>
<template><protein id="%1">%0</protein></template>
<t p1="ipi4355">casein</t>
<t p1="X33355">p53</t>

<template><disease id="%1">%0</disease></template>
<t p1="EFO_0000249">alzheimer disease</t>
<t p1="EFO_0003767">inflammatory bowel disease</t>
<t p1="EFO_0003767">IBD</t>
</mwt>
```
Running a dictionary-based tagger

```
> cat sample.txt | java -jar monq.jar monq.programs.DictFilter sample.mwt
```

output

The `<protein id="X33355">p53</protein>` homologue p73 accumulates in the nucleus and localizes to neurites and neurofibrillary tangles in `<disease id>Alzheimer disease</disease>` brain.

“The p53 homologue p73 accumulates in the nucleus and localizes to neurites and neurofibrillary tangles in Alzheimer disease brain.”
DictFilter Usage

usage: DictFilter -t type [-e elem] [-ie inEnc] [-oe outEnc] mwtfile
    -t: type of input: raw=plain ascii, xml=recognize and skip xml tags,
        elem=tag only within xml element given with option -e,
        1 string value from the list {raw, xml, elem}
    -e: specifies xml element within which to work, when '-t elem' is
        specified,
        1 string value,
        default: `plain`
    -ie: encoding used for input stream, guessed from input if not
        specified and then defaults to the platform encoding,
        1 string value
    -oe: encoding used for output stream, defaults to the platform encoding,
        1 string value
Session 3
Overview

• CTTV
  • Consortium of three Founding Members: GSK, EMBL-EBI and WTSI + Biogen (joined recently)
  • Integrates various evidences such as gene expression, variation and text mined data
  • A publicly available platform for target validation

• Europe PMC
  • Open access part:
    • 1 M OA full text
    • >25 M PubMed abstracts

• Goal: to push literature content from Europe PMC to the CTTV platform
Europe PMC CTTV literature component

- **Phase-I**: Pilot study:
  - Inflammatory Bowel Disease (IBD)
  - Sentence level target-IBD co-occurrences
  - Filtering out obvious false positives
  - Performance estimation

Disease => Disease/Phenotype
Europe PMC CTTV literature component (cont.)

- **Phase-II**: Extension:
  - 18 priority diseases (GSK focused)
  - Performance estimation

- **Phase-III**: Further Extension:
  - All 8.5K diseases
  - Disease name abbreviation filtering
  - Protein name abbreviation filtering
  - Negative sentence filtering
  - Article scoring
Dictionaries

- Core dictionaries: UniProt and EFO
- Dictionary Refinement
  - Screening names for false positives E.g. Common English words such as “Large”, HSC (Human Stem Cell)
  - Term length > 2
  - Greek letters and symbols
- Gene/Protein names
  - All UniProt (annotated): 393K
  - *Human UniProt (annotated): 100K terms
- Disease/Phenotype
  - EFO (disease/phenotypes): 26K terms
Filtering rules

- Round 1: No filtering
- Round 2:
  - UniProt – annotated human proteins (100K)
  - For documents in which “IBD” is the disease name, ensure that “inflammatory bowel disease” or similar also appears in the document (IBD is also used for “Infectious Bursal Disease”)
  - “Research” type articles only (full text)
  - Sections filtering e.g. remove Acknowledgements, Abbreviations, Methods
- Round 3:
  - Filtering out all the target-disease associations that occur only once in an article body (but not in Title/Abstract).

Results from filtering (IBD relevant only)

Number of publications from filtering

<table>
<thead>
<tr>
<th>Literature set</th>
<th>Number of publications</th>
</tr>
</thead>
<tbody>
<tr>
<td>round 1</td>
<td>8000</td>
</tr>
<tr>
<td>round 2</td>
<td>5000</td>
</tr>
<tr>
<td>round 3</td>
<td>3500</td>
</tr>
</tbody>
</table>

Number of unique targets from filtering

<table>
<thead>
<tr>
<th>Literature set</th>
<th>Number of unique targets</th>
</tr>
</thead>
<tbody>
<tr>
<td>round 1</td>
<td>3000</td>
</tr>
<tr>
<td>round 2</td>
<td>1300</td>
</tr>
<tr>
<td>round 3</td>
<td>900</td>
</tr>
</tbody>
</table>
Performance estimation for the pilot system

- 3 different IBD experts annotated 45 records each manually
  - Records => sentences identified by the system as target-IBD evidence

- Precision value of 79%

Precision = \( \frac{TP}{TP + FP} \)
Phase-II: Extending the study

- 18 different GSK focused priority diseases
  - Parkinson's Disease, Inflammatory Bowel Disease, Diabetes, Epilepsy, Cardiovascular Disease, etc.

- Performance estimation
  - Precision value of 78%
Phase –III: Further extension:

- All 8.5 K diseases (26K terms) from EFO
- Disease name abbreviation filtering
  - e.g. ALS (Amyotrophic Lateral Sclerosis) is ambiguous with “Advanced Life Support”
  - Potential abbreviation:
    - appears in the form of X….. Y…. Z…. (XYZ) AND
    - tagged as disease term, uppercase, length<6 AND
    - Long form from EFO exists in document OR Long form extracted from text contains any keywords such as disease, disorder, syndrome, defect, etc.
- Negative sentence filtering
  - Difficult
    - Only one clear evidence: "not associated"
Phase III – Cont...

• Protein name abbreviation filter

• e.g. GCP (Green cone photoreceptor pigment) is ambiguous with “Good Clinical Practice”

• Potential abbreviation:
  • appears in the form of X….. Y…. Z…. (XYZ) AND
  • tagged as protein name, uppercase, length<6 AND
  • (XYZ) appears more than 3 times in document body (full text only) OR
  • Long form matches with any term in Swissprot or Enzymes OR
  • Long form ends with (-ase/-ases) OR contains any of the keywords such as factor, receptor, gene, protein etc. OR
  • At least 4 sentences contain one of the keywords: “mutation, SNP, variation, gene” along with the protein name abbreviation candidate and a disease name

• Abbreviation filtering + negative sentence filtering : F-score of 92.3%
### Article Scoring

- **Abstract scoring:**

<table>
<thead>
<tr>
<th>Criteria</th>
<th>Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>First or second sentence</td>
<td>2</td>
</tr>
<tr>
<td>Last sentence</td>
<td>5</td>
</tr>
<tr>
<td>Any other sentences</td>
<td>3</td>
</tr>
<tr>
<td>Number of targets</td>
<td>0.2</td>
</tr>
<tr>
<td>Boosting up</td>
<td>Median value of article body scores from the open access set</td>
</tr>
</tbody>
</table>

- **Full text scoring:**

<table>
<thead>
<tr>
<th>Section</th>
<th>Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>TITLE</td>
<td>10</td>
</tr>
<tr>
<td>ABSTRACT</td>
<td>Same as above, except boosting up</td>
</tr>
<tr>
<td>RESULTS, FIG, TABLE</td>
<td>5</td>
</tr>
<tr>
<td>DISCUSS, CONCL</td>
<td>2</td>
</tr>
<tr>
<td>INTRO, CASE, APPENDIX, OTHER</td>
<td>1</td>
</tr>
</tbody>
</table>
Article Scoring Performance Estimation

- Mean Average Precision (MAP)
  - How well our scoring system performs by putting it into the concept of ranking
  - Takes into account the relative order of the documents retrieved by the system and give more weights to documents returned at higher ranks
  - Two different associations:
    - IGF1 and Diabetes
    - VMD2 and macular dystrophy
  - Top 25 documents are manually analysed
  - MAP is estimated as 81%
EuropePMC and other Public Data Sources in CTTV

<table>
<thead>
<tr>
<th>Evidence Type</th>
<th>Source</th>
<th>Target-Disease Records</th>
</tr>
</thead>
<tbody>
<tr>
<td>Marketed drugs</td>
<td>ChEMBL</td>
<td>20,169</td>
</tr>
<tr>
<td>Affected pathways</td>
<td>Reactome</td>
<td>9,002</td>
</tr>
<tr>
<td>Somatic mutations</td>
<td>Cancer Gene Census</td>
<td>12,079</td>
</tr>
<tr>
<td>Rare genetic mutations</td>
<td>Uniprot</td>
<td>9,119</td>
</tr>
<tr>
<td>Rare genetic mutations</td>
<td>European Variation Archive</td>
<td>19,096</td>
</tr>
<tr>
<td>Complex disease</td>
<td>GWAS catalog</td>
<td>26,860</td>
</tr>
<tr>
<td>Mouse models</td>
<td>Phenodigm</td>
<td>134,451</td>
</tr>
<tr>
<td>RNA expression</td>
<td>ArrayAtlas</td>
<td>300,256</td>
</tr>
<tr>
<td>Text mining</td>
<td>Europe PMC</td>
<td>3,176,463</td>
</tr>
</tbody>
</table>
### Statistics on the current literature data in CTTV

- **Top 3 disease/phenotypes:**

<table>
<thead>
<tr>
<th>Disease/Phenotype</th>
<th>#Records in the CTTV database</th>
<th>% in the CTTV database</th>
</tr>
</thead>
<tbody>
<tr>
<td>neoplasm</td>
<td>444938</td>
<td>14.35</td>
</tr>
<tr>
<td>cancer</td>
<td>175625</td>
<td>5.66</td>
</tr>
<tr>
<td>inflammation</td>
<td>139077</td>
<td>4.48</td>
</tr>
</tbody>
</table>

- **Top 3 targets:**

<table>
<thead>
<tr>
<th>Target</th>
<th>#Records in the CTTV database</th>
<th>% in the CTTV database</th>
</tr>
</thead>
<tbody>
<tr>
<td>TNF</td>
<td>94705</td>
<td>3.05</td>
</tr>
<tr>
<td>IL6</td>
<td>47022</td>
<td>1.52</td>
</tr>
<tr>
<td>EGFR</td>
<td>41402</td>
<td>1.33</td>
</tr>
</tbody>
</table>
Ongoing Work- EFO expansion

Overlapping triples (Gene, PMCID, Disease Term) between EFO, DO and UMLS

#Distinct UMLS terms=17,584
#Distinct DO terms=3,859
Next steps

- Continue with contributing to EFO expansion (ontologies to consider: ORDO, MP, HPO, etc.)

- Extending the study to identify:
  - Mechanism type
    - (inhibit, activate, upregulate downregulate, etc)
  - Article type
    - (genetic, clinical trial, pre-clinical trial)
  - Experiment
    - (cell line, organism (human/other?))
Europe PMC CTTV component

**CYSLTR1 associations with memory impairment**

**CYSLTR1**

- **cysteinyl leukotriene receptor 1**
- **Synonyms:** CysLT1, CysLT1...

Receptor for cysteinyl leukotrienes mediating bronchoconstriction of individuals with and without asthma. Stimulation by LTC4 results in the contraction and proliferation of smooth muscle, edema, etc...

**memory impairment**

- **Synonyms:** Memory Impairment, Adverse Event...

---

<table>
<thead>
<tr>
<th>Genetic associations</th>
<th>Somatic mutations</th>
<th>Known drugs</th>
<th>RNA expression</th>
<th>Affected pathways</th>
</tr>
</thead>
</table>

---

**Text mining**

Source: DisGeNET and Europe PMC

Showing 1 to 1 of 1 entries

<table>
<thead>
<tr>
<th>Disease</th>
<th>Publication</th>
<th>Matched sentences</th>
<th>Year</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>memory impairment</td>
<td>Montelukast targeting the cysteinyl leukotriene receptor 1 ameliorates Aβ1-42-induced memory impairment and neuroinflammatory and apoptotic responses in mice. Li S et al. Neuropharmacology.</td>
<td></td>
<td>2014</td>
<td>Europe PMC</td>
</tr>
</tbody>
</table>

---

Mouse models

- [https://www.targetvalidation.org](https://www.targetvalidation.org)
Contacts

Şenay Kafkas (kafkas@ebi.ac.uk)

Jee-Hyub Kim (jhkim@ebi.ac.uk)