Retrieving Biomedical Literature - An Open Source Search Engine Based on Open Access Resources

Hayda Almeida, Ludovic Jean-Louis, Marie-Jean Meurs

CONCORDIA GENOMICS

UQÀM

Biocuration 2016, Genève

April 2016

Biomedical Literature Retrieval

- Scientific databases → support for research and health care
- Large amount of open access data available

PubMed BD PMC OA

$24{,}000{,}000^{+}$ + $1{,}200{,}000^{+}$ = **25,403,053 records**

PY: Since 1809 Since 1973

- Retrieval of relevant information → critical task
- Scientific journal articles → input for many tasks (Almeida et al., 2014)

Scientific Database Search: Challenges

1 Article content searched

PubMed — article abstract

PMC — article full-text

Use of full-text search:

- Better support for literature analysis tasks (Gay et al., 2005)
- Improvement in search results (Nourbakhsh et al., 2012)
- Access to more relevant information in articles (Van Auken et al., 2014)

Scientific Database Search: Challenges

2 Express search in query language

- Users frequently reformulate queries (Dogan et al., 2009)
- Few users generate advanced queries (Shariff et al., 2013)
- Most searches made by inexperienced users (Yoo and Mosa, 2015)

Natural language: `alpha-amylase from Cryptococcus flavus`

Query language: `alpha-amylase OR alpha amylase AND (cryptococcus OR (cryptococcus AND flavus))`

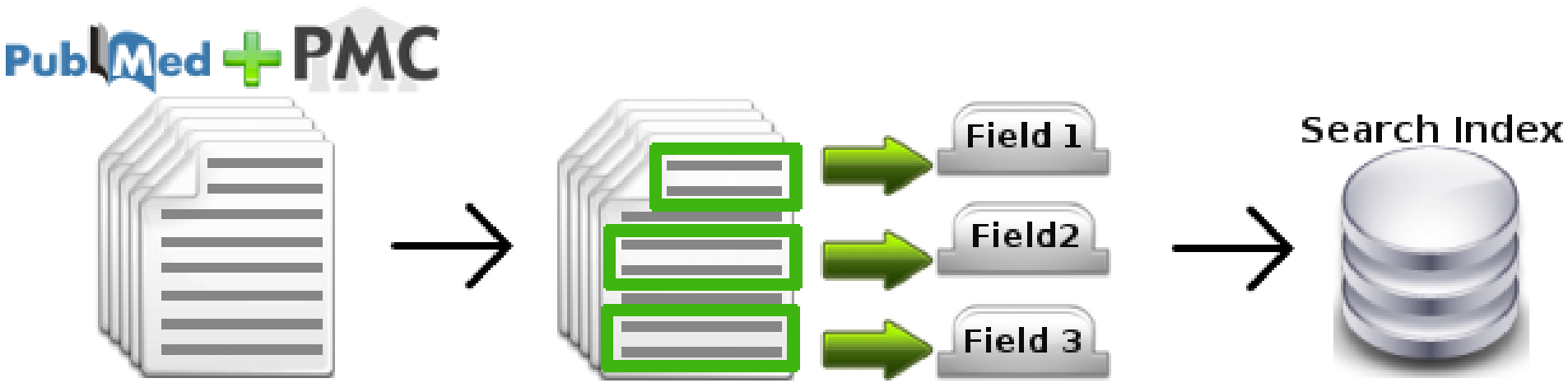
Suggested Approach

- Search engine for biomedical open access data
- Ability to handle natural language queries
- Document Indexing Module

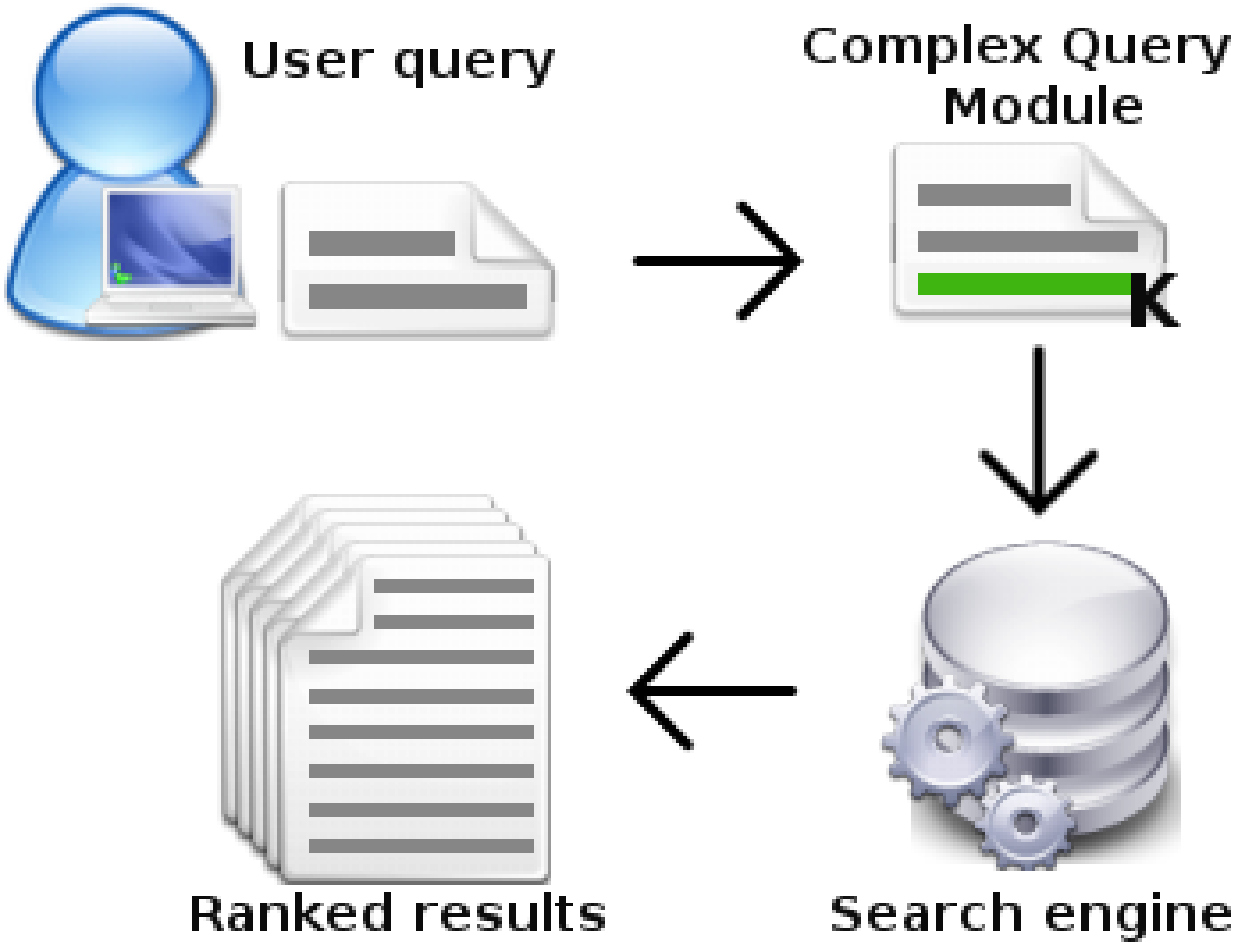
  Parsing $\rightarrow$ XML (PubMed) and NXML (PMC)
  Relevant fields $\rightarrow$ full-text, abstract, metadata
  Indexing $\rightarrow$ map fields to index schema
- Complex Query Module

  User input $\rightarrow$ natural language processing
  Query types & query strategies per type
  Query expansion $\rightarrow$ UMLS Metathesaurus annotations

Pipeline: Document Indexing



Pipeline: Query Search



Document Indexing Module

- Based on Apache Solr
- One index entry per document
- Document semantic representation: {document field, content}

Field	PubMed BD	PMC OA	Field	PubMed BD	PMC OA
1 Article title	✓	✓	**8** Reference title	✗	✓
2 Journal title	✓	✓	**9** Reference IDs	✗	✓
3 Abstract	✓	✓	**10** Object captions	✗	✓
4 Body section titles	✗	✓	**11** PMCID	✓	✓
5 Body full content	✗	✓	**12** PMID	✓	✓
6 Author names	✓	✓	**13** Article keywords	✗	✓
7 Reference authors	✓	✓	**14** Publication year	✓	✓

Complex Query Module: Query Types

- Keyword query, K_Q

  No stop-words among query terms

  `"AIDS versus HIV"`

- Open Question query, O_Q

  Presents interrogative cues

  `"what is the difference between HIV and AIDS?"`

- Statement query, S_Q

  Does not present interrogative cues
  Has stop-words

  `"the difference between HIV and AIDS"`

Complex Query Module: Query Generation

- Each query type $\rightarrow$ different strategy
- Search Fields $\rightarrow$ where to look for query terms
- Phrase Search Fields $\rightarrow$ look for query terms appearing in sequence
- Boost $\rightarrow$ increase document relevance with a coefficient in query time

Type	Search Fields	Boost?	Phrase Search Fields	Boost?
K_Q	abstract, body, keywords {...}	abstract, body	title, body, authors {...}	title, authors
O_Q	title, abstract, body {...}	captions, body {...}	captions, abstract {...}	abstract, body
S_Q	body, authors, keywords {...}	title, captions {...}	title, abstract {...}	title, body

Complex Query Module: Query Expansion

- MetaMap (Aronson and Lang, 2010) $\rightarrow$ UMLS Methathesaurus
- Avoid redundancy $\rightarrow$ annotations without any terms in user query

User query: `"AIDS versus HIV"`

MetaMap annotations:

```
"HIV+ [HIV Seropositivity]"
"AIDS [Acquired Immunodeficiency Syndrome]"
"HIV [HIV]"
```

Expanded query:

`"AIDS versus HIV Acquired Immunodeficiency Syndrome"`

Preliminary Evaluation Data

- Large data $\rightarrow$ challenge finding manual annotations
- 19 manually annotated sets {`query, target article ID`}
  - Biocurators support: mycoCLAP (Strasser et al., 2015) database
  - Each enzyme entry $\rightarrow$ 1^{+} article(s)
  - Articles retrieved from scientific literature databases
- 9 {`query, PMCID`}, 10 {`query, PMID`}

$Q_{\#}$	**Target article ID**	**User query**	**mycoCLAP ID**
Q_3	PMC2780388	characterization of GH5 beta-mannanase enzyme from Aspergillus niger	MAN5A_ASPNG
Q_4	PMC3092853	characterization of GH16 beta-glucanase from Aspergillus fumigatus	MLG16B_ASPFU
Q_{15}	PMID1400249	characterization of Candida albicans maltase	AGL13B_CANAL
Q_{16}	PMID12761390	beta-1,4-galactanases from Humicola insolens and Myceliophthora thermophila	GAN53A_HUMIN

Evaluation Metrics

- Pseudo-judgement $\rightarrow$ top 20 ranked results
- Reciprocal Rank (RR)

  Computed for each query

  Inverse of target article ranking

$$\mathrm{RR} = \frac{1}{position}$$

- Mean Reciprocal Rank (MRR)

  Computed for all queries

  RR average for the 19 {`query, target article ID`} sets

$$\mathrm{MRR} = \frac{1}{|Q|} \sum_{i=1}^{|Q|} \frac{1}{position}$$

Preliminary Results

$Q_{\#}$	PMC rank	bioMine rank	bioMine RR score	$Q_{\#}$	PubMed rank	bioMine rank	bioMine RR score
Q_1	3	**2**	0.500	Q_{10}	2	**1**	1.000
Q_2	1	20	0.050	Q_{11}	N/A	**7**	0.143
Q_3	1	2	0.500	Q_{12}	1	**1**	1.000
Q_4	2	8	0.125	Q_{13}	2	**1**	1.000
Q_5	2	13	0.077	Q_{14}	1	**1**	1.000
Q_6	9	**1**	1.000	Q_{15}	2	**1**	1.000
Q_7	2	5	0.200	Q_{16}	1	N/A	0.000
Q_8	1	17	0.059	Q_{17}	N/A	**1**	1.000
Q_9	1	10	0.100	Q_{18}	1	N/A	0.000
				Q_{19}	1	**1**	1.000
total # of queries = 19				**MRR = 0.513**			

Preliminary Results

$Q_{\#}$	**PMC rank**	**bioMine rank**	**bioMine RR score**	$Q_{\#}$	**PubMed rank**	**bioMine rank**	**bioMine RR score**
Q_1	3	2	0.500	Q_{10}	2	1	1.000
Q_2	1	20	0.050	Q_{11}	N/A	7	0.143
Q_3	1	2	0.500	Q_{12}	1	1	1.000
Q_4	2	8	0.125	Q_{13}	2	1	1.000
Q_5	2	13	0.077	Q_{14}	1	1	1.000
Q_6	9	1	1.000	Q_{15}	2	1	1.000
Q_7	2	5	0.200	Q_{16}	1	N/A	0.000
Q_8	1	17	0.059	Q_{17}	N/A	1	1.000
Q_9	1	10	0.100	Q_{18}	1	N/A	0.000
				Q_{19}	1	1	1.000
total # of queries = 19				**MRR = 0.513**			

Preliminary Results

$Q_{\#}$	PMC rank	bioMine rank	bioMine RR score	$Q_{\#}$	PubMed rank	bioMine rank	bioMine RR score
Q_1	3	2	0.500	Q_{10}	2	1	1.000
Q_2	1	20	0.050	Q_{11}	N/A	7	0.143
Q_3	1	2	0.500	Q_{12}	1	1	1.000
Q_4	2	8	0.125	Q_{13}	2	1	1.000
Q_5	2	13	0.077	Q_{14}	1	1	1.000
Q_6	9	1	1.000	Q_{15}	2	1	1.000
Q_7	2	5	0.200	Q_{16}	1	N/A	0.000
Q_8	1	17	0.059	Q_{17}	N/A	1	1.000
Q_9	1	10	0.100	Q_{18}	1	N/A	0.000
				Q_{19}	1	1	1.000
total # of queries = 19				**MRR = 0.513**			

Preliminary Results

$Q_{\#}$	**PMC rank**	**bioMine rank**	**bioMine RR score**	$Q_{\#}$	**PubMed rank**	**bioMine rank**	**bioMine RR score**
Q_1	3	2	0.500	Q_{10}	2	1	1.000
Q_2	1	20	0.050	Q_{11}	N/A	7	0.143
Q_3	1	2	0.500	Q_{12}	1	1	1.000
Q_4	2	8	0.125	Q_{13}	2	1	1.000
Q_5	2	13	0.077	Q_{14}	1	1	1.000
Q_6	9	1	1.000	Q_{15}	2	1	1.000
Q_7	2	5	0.200	Q_{16}	1	N/A	0.000
Q_8	1	17	0.059	Q_{17}	N/A	1	1.000
Q_9	1	10	0.100	Q_{18}	1	N/A	0.000
				Q_{19}	1	1	1.000
total # of queries = 19				**MRR = 0.513**			

Preliminary Results

$Q_{\#}$	**PMC rank**	**bioMine rank**	**bioMine RR score**	$Q_{\#}$	**PubMed rank**	**bioMine rank**	**bioMine RR score**
Q_1	3	2	0.500	Q_{10}	2	1	1.000
Q_2	1	20	0.050	Q_{11}	N/A	7	0.143
Q_3	1	2	0.500	Q_{12}	1	1	1.000
Q_4	2	8	0.125	Q_{13}	2	1	1.000
Q_5	2	13	0.077	Q_{14}	1	1	1.000
Q_6	9	1	1.000	Q_{15}	2	1	1.000
Q_7	2	5	0.200	Q_{16}	1	N/A	0.000
Q_8	1	17	0.059	Q_{17}	N/A	1	1.000
Q_9	1	10	0.100	Q_{18}	1	N/A	0.000
				Q_{19}	1	1	1.000
total # of queries = 19				**MRR = 0.513**			

Conclusion and Ongoing Work

- Scientific literature search in article abstracts and full-text
- Processing of natural language queries
- Target articles ranked in bioMine at first position ≈50% of the time
- Use of open access data
- Source code publicly available

`https://github.com/BigMiners/bioMine`

Next steps

- Improvement of full-text document retrieval
- Development of web-based user interface

Thank you!

Questions?

References

Almeida et al., *Machine Learning for Biomedical Literature Triage*, PLOS ONE, 2014.

Gay et al., *Semi-automatic Indexing of Full Text Biomedical Articles*, AMIA Annual Symposium Proceedings, 2005.

Nourbakhsh E. et al., *Medical Literature Searches: A Comparison of PubMed and Google Scholar*, Health Information & Libraries Journal, 2012.

Van Auken et al., *BC4GO: A Full-text Corpus for the BioCreative IV GO Task*, Database, 2014.

Dogan et al., *Understanding PubMed User Search Behaviour through Log Analysis*, Database, 2009.

Shariff et al., *Retrieving Clinical Evidence: A Comparison of PubMed and Google Scholar for Quick Clinical Searches, Journal of Medical Internet Research*, 2013.

Yoo and Mosa, *Analysis of PubMed User Sessions Using a Full-Day PubMed Query Log: A Comparison of Experienced and Nonexperienced PubMed Users*, Journal of Medical Internet Research, 2015.

Aronson A. and Lang F., *An Overview of MetaMap: Historical Perspective and Recent Advances*, Journal of the American Medical Informatics Association, 2010.

Strasser K. et al., *mycoCLAP, the Database for Characterized Lignocellulose-active Proteins of Fungal Origin: Resource and Text Mining Curation Support*, Database, 2015.

Corpus Description

Baseline Database (BD) Files

- Journal article abstract, citations, books
- Publication years since at least 1809
- 24,350,000$^+$ entries

Open Access (OA) Subset

- Full-text journal articles
- Publication years since at least 1973
- 1,200,000$^+$ entries

Total entries indexed: **25,403,053**

Evaluation Data: `query, PMCID`

$Q_{\#}$	**Target article ID**	**User query**	**mycoCLAP ID**
Q_1	PMC3068306	alpha-amylase from Cryptococcus flavus activity characterization	AMY13A_CRYFL
Q_2	PMC3312866	Aspergillus fumigatus beta-glucosidase purification and characterization	BGL3C_ASPFU
Q_3	PMC2780388	characterization of GH5 beta-mannanase enzyme from Aspergillus niger	MAN5A_ASPNG
Q_4	PMC3092853	characterization of GH16 beta-glucanase from Aspergillus fumigatus	MLG16B_ASPFU
Q_5	PMC3180650	purification and characterization of an exo-polygalacturonase from Fusarium oxysporum	PGX28B_FUSOX
Q_6	PMC3223205	Phanerochaete chrysosporium GH61 purification and characterization	PMO9D_PHACH
Q_7	PMC3312857	purification and characterization of an alpha-L-rhamnosidase from Aspergillus nidulans	RHA78E_EMENI
Q_8	PMC2291056	xylanase characterization from Leucoagaricus gongylophorus	XYN11A_LEUGO
Q_9	PMC2702311	recombinant expression and characterization of xylanase from Trichoderma reesei	XYN11B_TRIRE

Evaluation Data: `query, PMID`

$Q_{\#}$	Target article ID	User query	mycoCLAP ID
Q_{10}	PMID20562284	bifunctional alpha-L-arabinofuranosidase /xylobiohydrolase from Penicillium purpurogenum	ZAX43C_PENPU
Q_{11}	PMID10215597	enzymatic properties alpha-mannosidase Aspergillus saitoi	MSD47S_ASPPH
Q_{12}	PMID20709852	characterization of Magnaporthe oryzae cellobiohydrolase	CBH6A_MAGOR
Q_{13}	PMID9758835	substrate specificity of alpha-L-arabinofuranosidase from Aspergillus awamori	ABF51A_ASPAW
Q_{14}	PMID7708682	cloning and characterization Candida albicans chitinase	CHI18B_CANAL
Q_{15}	PMID1400249	characterization of Candida albicans maltase	AGL13B_CANAL
Q_{16}	PMID12761390	beta-1,4-galactanases from Humicola insolens and Myceliophthora thermophila	GAN53A_HUMIN
Q_{17}	PMID12427996	Neotyphodium sp beta-1,6-glucanase expression and characterization	BGN5A_NEOSP
Q_{18}	PMID21653698	purification of endo-beta-1,3-galactanase from Flammulina velutipes	EBG16A_FLAVE
Q_{19}	PMID9872754	Aspergillus oryzae beta-xylosidase optimum pH and temperature	XYL3A_ASPOR