


# Text Mining for Biomedicine

**He Tan**






Laboratory for Intelligent Information Systems  
Institutionen för datavetenskap



1

## Why?

- Individual gene study → large scale analysis
- Genomics Biology: increasing number of genomes, sequences, proteins
- Difficulty interpreting large scale experimental results, e.g. Y2H, microarrays, etc.

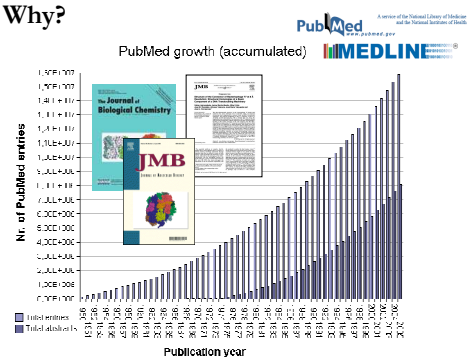






Department of Computer and Information Science (IDA)  
Linköpings universitet, Sweden

2

## Why?

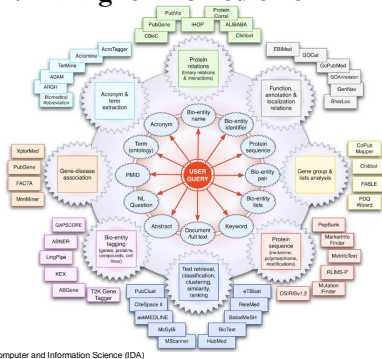
PubMed growth (accumulated)



Department of Computer and Information Science (IDA)  
Linköpings universitet, Sweden

3

## Text Mining for Biomedicine



Department of Computer and Information Science (IDA)  
Linköpings universitet, Sweden

4


## Disciplines

- Information retrieval**
  - finding the papers
- Entity recognition**
  - identifying the entities
- Information extraction**
  - formalizing the facts
- Text mining**
  - finding nuggets in the literature
- Integration**
  - combining text and biological data

Department of Computer and Information Science (IDA)  
Linköpings universitet, Sweden

5

## Status



Department of Computer and Information Science (IDA)  
Linköpings universitet, Sweden

Literature mining for the biologist:  
from information retrieval to biological discovery  
Jensen et al., *Nature Reviews Genetics*, 2006

6

## Challenges

- KDD Cup 2002
- TREC Genomics Tracks 2003 - 2007
- BioNLP/JNLPBA 2004
- LLL 2005
- BioCreative I (2004) & II (2006)

Department of Computer and Information Science (IDA)  
Linköpings universitet, Sweden

7

## Biomedical Language

- Heavy use of domain specific terminology
  - e.g. *chemoattractant, fibroblasts, angiogenesis*
- No standard nomenclatures supported by scientific community
  - HUGO Gene Nomenclature ?
- Rapid growth of new names and new senses
  - 'This disorder maps to chromosome 7q11-21, and this locus was named *CLAM*. '[PMID:12771259 ]
- Most words with low frequency (data sparseness)

Department of Computer and Information Science (IDA)  
Linköpings universitet, Sweden

8

## Biomedical Language

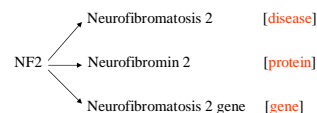
- short forms and abbreviations are often used
  - e.g. *vascular endothelial growth factor (VEGF)*
- genes have often synonyms
  - e.g. *Thermoactinomyces candidus*  
vs. *Thermoactinomyces vulgaris*
- Orthographic variants
  - e.g. *TNF $\alpha$* , *TNF-alpha* and *TNF alpha* (without hyphen)

Department of Computer and Information Science (IDA)  
Linköpings universitet, Sweden

9

## Biomedical Language

- A name
  - may refer to a particular gene
  - may include homologues of this gene in other organisms
  - may denote an RNA, DNA, or the protein the gene encodes
  - may be restricted to a specific splice variant



Department of Computer and Information Science (IDA)  
Linköpings universitet, Sweden

10

## Text Mining and Natural language processing (NLP)

- NLP
  - techniques to analyze, understand and generate natural language used by human, e.g. free text
- Text mining
  - generally relies on NLP in e.g. IR, NER, IE

Department of Computer and Information Science (IDA)  
Linköpings universitet, Sweden

11

## The framework of NLP

- Tokenization
  - text  $\rightarrow$  tokens
- Lemmatization
  - word  $\rightarrow$  its lemma
- POS tagging
  - e.g. NP-VP-NP
- B-I-O tagging
  - define the begin and the end of the name entity
- **Semantics** interpretation
- **Pragmatics** analysis

Department of Computer and Information Science (IDA)  
Linköpings universitet, Sweden

12



## GENIA corpus

- POS annotation
- Trebank
- Coreference annotation
- Term annotation
- Event annotation
- Cellular localization
- Disease-Gene association
- Pathway corpus

Department of Computer and Information Science (IDA)  
Linköpings universitet, Sweden

19

## GENIA corpus

Department of Computer and Information Science (IDA)  
Linköpings universitet, Sweden

20

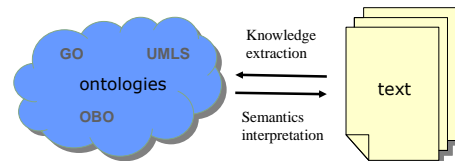
## Acronyms

- Many resources available
  - AcroMine  
<http://www.nactem.ac.uk/software/acromine>
  - ARGH: Biomedical Acronym  
<http://lethargy.swmed.edu/ARGH/argh.asp>
  - Stanford Biomedical Abbreviation Server  
<http://bionlp.stanford.edu/abbreviation/>
  - AcroMed  
<http://medstract.med.tufts.edu/acro1.1/index.htm>
  - SaRAD  
<http://www.hpl.hp.com/research/idl/projects/abbrev.html>

Department of Computer and Information Science (IDA)  
Linköpings universitet, Sweden

21

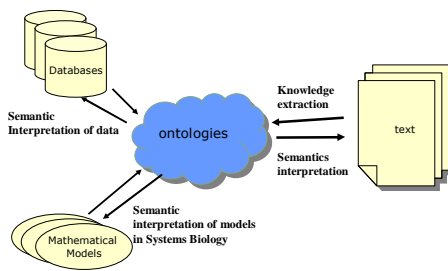
## Text Mining and Ontologies



Department of Computer and Information Science (IDA)  
Linköpings universitet, Sweden

22

## Text Mining and Ontologies



Department of Computer and Information Science (IDA)  
Linköpings universitet, Sweden

23

## The Gene Ontology

- Controlled vocabulary for the annotation of gene products

26292 terms, 98.4% with definitions

15632 biological\_process

2233 cellular\_component

8427 molecular\_function

November 10, 2008 at 2:00 Pacific time

five types of relationships:

*is\_a, part\_of,*

*regulates,*

*positively\_regulates and negatively\_regulates.*

Department of Computer and Information Science (IDA)  
Linköpings universitet, Sweden

24





## Ad hoc IR systems

- Query is typically,
  - Boolean model
    - e.g. *yeast AND cell cycle*
- By automatically expanding queries with additional search terms, recall can be improved
  - Stemming removes common endings (*yeast / yeasts*)
  - Stop word removal, such as *the* and *an*
  - Case folding (*yeast / YEAST*)
  - Thesauri can be used to expand queries with synonyms and/or abbreviations (*yeast / S. cerevisiae*)

Department of Computer and Information Science (IDA)  
Linköpings universitet, Sweden

37

## Ad hoc IR systems

Department of Computer and Information Science (IDA)  
Linköpings universitet, Sweden

38

## Ad hoc IR systems

Department of Computer and Information Science (IDA)  
Linköpings universitet, Sweden

39

## Document Similarity

- The similarity of two documents can be defined based on their word content
  - Each document is represented in a vector of word.
  - Word is weighted according to their frequency within the document and the document collection.
  - The similarity is calculated based on those vectors

Department of Computer and Information Science (IDA)  
Linköpings universitet, Sweden

40

## Document Clustering

- Unsupervised clustering algorithms
  - All pairwise document similarities are calculated
  - Clusters of "similar documents"
- In ad hoc IR systems
  - Rather than matching the query against each document only, the N most similar documents are also considered

Department of Computer and Information Science (IDA)  
Linköpings universitet, Sweden

41

## Document Clustering

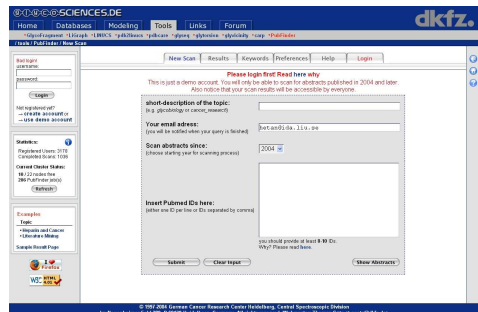
Department of Computer and Information Science (IDA)  
Linköpings universitet, Sweden

42

## Document Classification

- Statistical machine learning methods
    - A pre-defined set of document classes
    - Each class is defined by manually assigning a number of documents to it
- classified documents

## Document Classification



## Information Retrieval

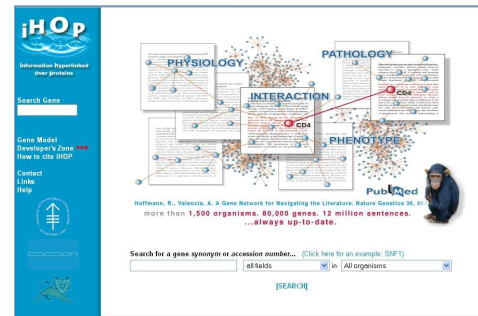
- The art is to find the relevant papers even if they do not actually match the query

Biological database:  
Organism: [ Saccharomyces cerevisiae ]  
GO process: cell cycle

Mitotic cyclin (Cln2)-bound Cdc28 (Cdk1 homolog) directly phosphorylated Swe1 and this modification served as a priming step to promote subsequent Cdc5-dependent Swe1 hyperphosphorylation and degradation

- The next logical step is to use ontologies to make complex inferences (yeast cell cycle / Cdc28)

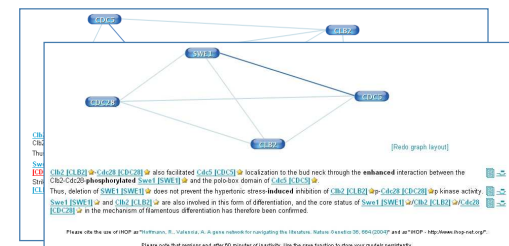
## iHOP (information Hyperlinked over Proteins)



## iHOP



## iHOP





# Wikiprofessional

Department of Computer and Information Science (IDA)  
Linköpings universitet, Sweden

49

# WikiProfessional

Department of Computer and Information Science (IDA)  
Linköpings universitet, Sweden

50

# Opinions from leading scientists

- Fusing literature and biological databases through text mining
- Interactivity and user interfaces
- Tool integration
- Text mining resources

Text mining for biology - the way forward: opinions from leading scientists  
Russ B Altman, et al. *Genome Biology* 2008, 9(Suppl 2):S7

Department of Computer and Information Science (IDA)  
Linköpings universitet, Sweden

51