Knowledge Extraction and Text Mining in Bioinformatics

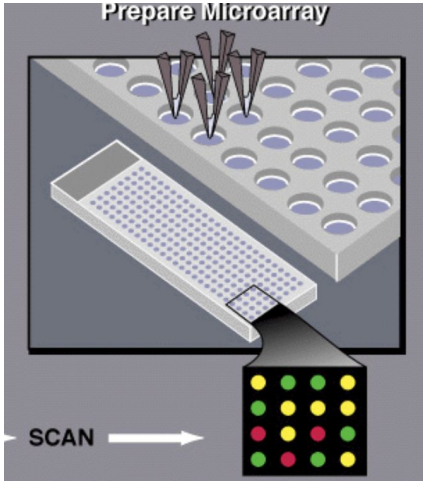
Presented By
Anmol Sharma
MSc Student in Computing Science
Medical Image Analysis Laboratory
Simon Fraser University
Burnaby, BC Canada.
20 September 2018

Overview

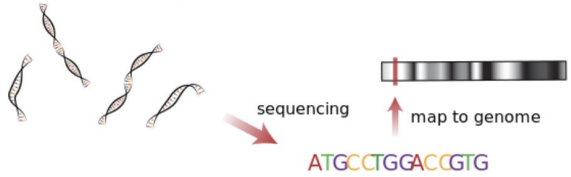
- Problem Statement
- Background
  - A Perspective on Bioinformatics Data
  - Text Mining and Natural Language Processing (NLP)
- Application of Text Mining in Biomedical Text Analysis
  - Biomedical Entity Tagging
  - Information Retrieval
  - Extracting Protein-Protein/Gene-Gene Interactions in Free Text
  - Extracting Gene-Trait Interactions
  - Extracting Train-Gene Interactions
- Application of Text Mining in DNA/RNA Sequencing
  - Sequence Alignment in DNA/RNA
- Remarks
- Future Work
- References

Problem Statement

- **Drug target search**
  - Relies heavily on inferential computational methods working with sequence and structure similarities.
  - Commonly followed by time consuming manual examination
- **Automated curation of large data streams**
  - Micro-arrays, two-hybrid systems provide high throughput large data collections.
  - High data volume makes interpretation challenging.
- **Protein Function Prediction**
  - Predicted by extrapolation of information gathered from small set of proteins.
  - Performed by experts and can have inter-observer variance.
- **Link unstructured data to structured annotation databases**
  - How to link free-flowing text data to structured databases automatically?
- **Gene Ontology**
  - Manual annotation of proteins from model organisms.



PubMed.gov
US National Library of Medicine
National Institutes of Health



A Perspective on Bioinformatics Data

- **Bioinformaticians handle a broad spectrum of heterogeneous data.**
- **A vast collection of databases with keywords, protein sequences, abstracts and structural information.**
  - INTACT [Hermjakob et al., 2004], BIND [Bader et al., 2003], DIP [Xenarios et al., 2002]
- **Free textual data**
  - 12 million abstracts in national Library of Medicine (NLM) [Wheeler et al., 2003]
- **Medical Ontologies**
  - Unified Medical Language Systems (UMLS) metathesaurus [Bodenreider, 2004]
  - Contains 2 million names for 0.9 million concepts.
- **Availability of this data is a required precursor to any text mining solution**

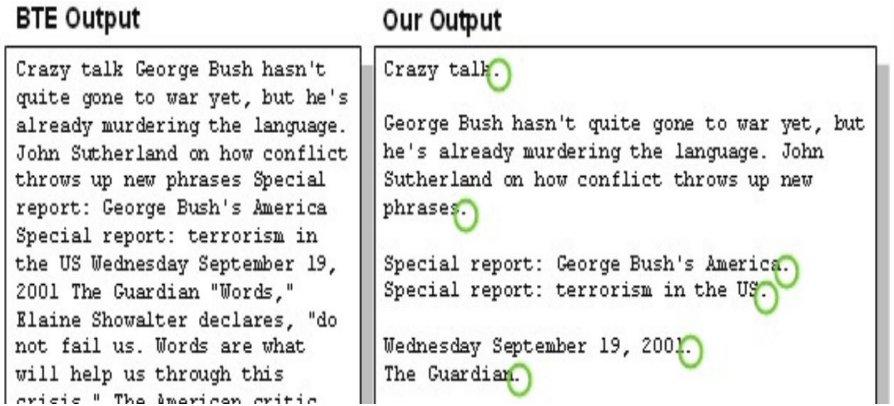
NIH U.S. National Library of Medicine

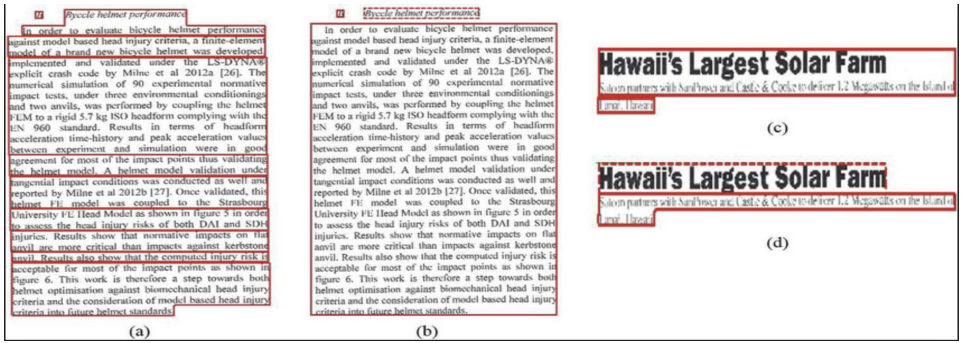
Unified Medical Language System® (UMLS®)

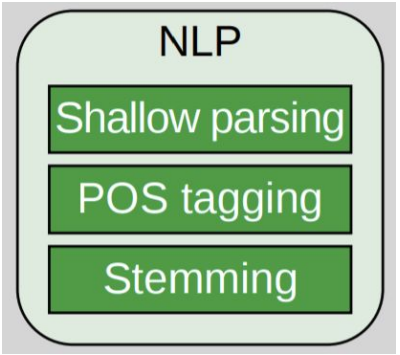
EMBL-EBI IntAct

Text Mining and Natural Language Processing (NLP)

- **Text Mining and Natural language** processing (NLP) is a subfield of computer science
- Concerned with the **interactions between computers and human (natural) languages.**
- In particular, how to program computers to process and analyze large amounts of natural language data.







Applications of Text Mining in Biomedical Text Analysis

Biological Entity Tagging (Named Entity Recognition)

- **Recognition and tagging of salient entities** in an **unstructured text** is of great importance for searching, curation, and **to identify relationship with other available data**.
- **Example:**
  - "*erythroid myeloid ontogenic factor, changed to Zbtb7 after a lawsuit was threatened*"
  - What are the entities of interest here?
  - What is their relationship?
- **Query unstructured text, returns relevant documents.**
- **Challenged by**
  - Complex biomedical language
  - Ambiguous names (proteins may be referred to by their name or symbol)
  - Ambiguous meaning (same name for a gene and experiment).

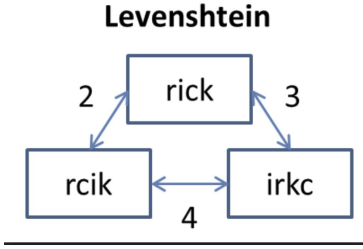
Jim bought 300 shares of Acme Corp. in 2006.

$[\text{Jim}]_{\text{Person}}$ bought 300 shares of $[\text{Acme Corp.}]_{\text{Organization}}$ in $[\text{2006}]_{\text{Time}}$.

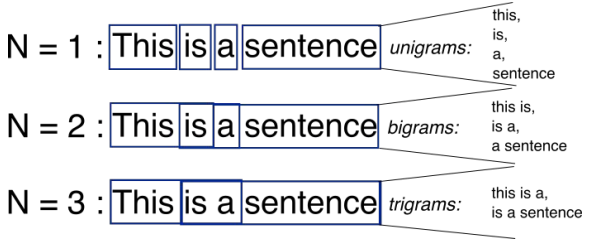
"Thus, $\text{CIITA}_{PROTEIN}$ not only activates the expression of $\text{class II genes}_{DNA}$ but recruits another B cell-specific coactivator to increase transcriptional activity of $\text{class II promoters}_{DNA}$ in $\text{B cells}_{CELLTYPE}$."

Biological Entity Tagging (Named Entity Recognition)

- **Different approaches have been adopted to address this problem.**
- **Ad-hoc rule-based approaches**
  - [Fukuda et al., 1998] Maintain a dictionary of words, and use exact or fuzzy matching
  - [Krauthammer et al., 2000] using Lavenshtein Distance [Navarro, 2001]
- **Machine Learning based approaches (pre-deep learning)**
  - Training an ML classifier using hand-crafted text features like POS tags, bigram/trigram frequencies , suffixes [Tanabe and Wilbur, 2002]
  - Support Vector Machine [Kazama et al., 2002] with hand-crafted binary features for named entity recognition
  - Bayesian Learning, decision trees and inductive rule learning [Hatzivassiloglou et al., 2001].
- **Deep Learning (State-of-art)**
  - Recently, state of art deep learning based approaches (LSTM/RNN) have been applied for biomedical NER [Wang et al., 2018, Dang et al., 2018]



POS tag	Tag Name	Relative Frequency (%)
N	Noun	39.74
PREP	Preposition	11.25
PUNC	Punctuation	10.27
AJ	Adjective	9.27
V	Verb	8.89
CON	Conjunction	8.48
NUM	Number	3.13
PRO	Pronoun	2.58
DET	Determiner	2.50
ADV	Adverb	1.84
POSTP	Postposition	1.47
RES	Residual	0.37
CL	Classifier	0.21
INT	Interjection	0.01



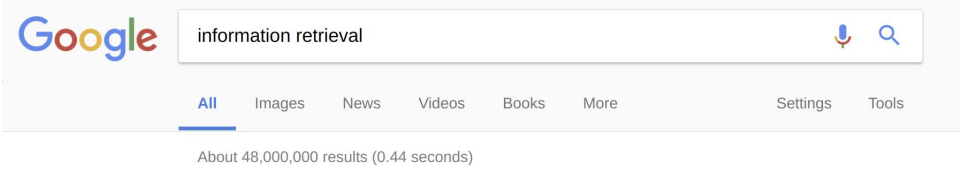
$$w_{k,i} = \begin{cases} 1 & \text{if a word at } k, W_k, \text{ is the } i\text{th word in the vocabulary } \mathcal{V} \\ 0 & \text{otherwise} \quad (\textit{word feature}) \end{cases}$$

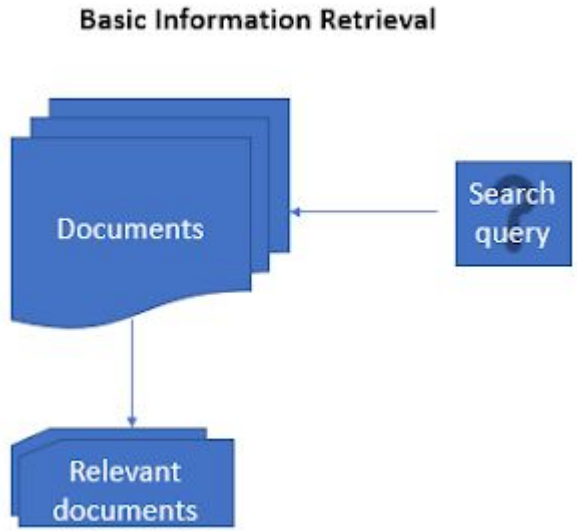
$$pos_{k,i} = \begin{cases} 1 & \text{if } W_k \text{ is assigned the } i\text{th POS tag in the POS tag list } \mathcal{POS} \\ 0 & \text{otherwise} \quad (\textit{part-of-speech feature}) \end{cases}$$

$$pre_{k,i} = \begin{cases} 1 & \text{if } W_k \text{ starts with the } i\text{th prefix in the prefix list } \mathcal{P} \\ 0 & \text{otherwise} \quad (\textit{prefix feature}) \end{cases}$$

Information Retrieval (IR)

- **Information Retrieval (IR)** recovers **relevant information given a free text query.**
- Information retrieval is the science of searching for information in a document or database, and also searching for metadata that describe data.
    - Data can be
        - Text
        - Images
        - Media Files





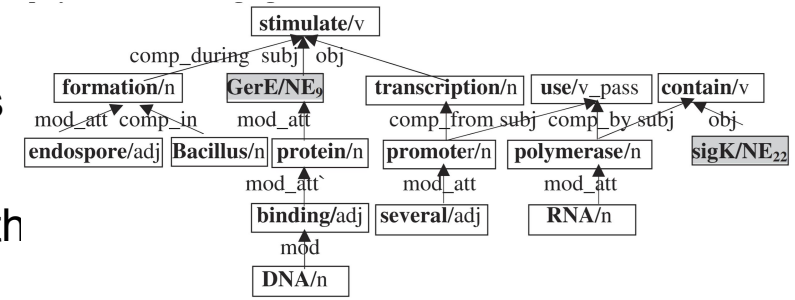
Information Retrieval (IR)

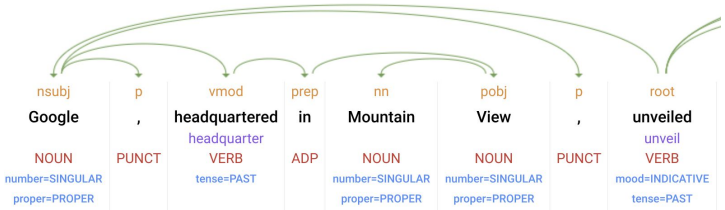
- Techniques such as "**document neighboring**" were first applied in [Wilbur and Coffee, 1994].
- Idea is to **keep relevant, connected documents close to each other in a code-word space**.
- The idea has been built upon recently, using **word2vec** [Mikolov et al., 2013], **doc2vec** [Le and Mikolov, 2014], **sentence2vec** [Le and Mikolov, 2014], which assigns code-words which are close to each other in vector space, and can be used to train ML classifiers.
- The current state-of-art in biomedical NER uses the abpve code-word representation of the words
  - One such method was proposed in [Wang et al., 2018] which uses word2vec embeddings to represent biomedical text.

Type of relationship	Word Pair 1		Word Pair 2	
Common capital city	Athens	Greece	Oslo	Norway
All capital cities	Astana	Kazakhstan	Harare	Zimbabwe
Currency	Angola	kwanza	Iran	rial
City-in-state	Chicago	Illinois	Stockton	California
Man-Woman	brother	sister	grandson	granddaughter
Adjective to adverb	apparent	apparently	rapid	rapidly
Opposite	possibly	impossibly	ethical	unethical
Comparative	great	greater	tough	tougher
Superlative	easy	easiest	lucky	luckiest
Present Participle	think	thinking	read	reading
Nationality adjective	Switzerland	Swiss	Cambodia	Cambodian
Past tense	walking	walked	swimming	swam
Plural nouns	mouse	mice	dollar	dollars
Plural verbs	work	works	speak	speaks

Extracting Protein-Protein/Gene-Gene Interactions

- Extracting the protein interaction relationship is an important research topic in bioinformatics and systems biology.
- Extracting protein-protein interactions involves searching for two proteins in the text and determining whether they interact with each oth
- In previous studies, researchers searched for protein-protein interactions manually.
- Challenging because:
  - No unified naming rule for proteins has been established yet.
  - Many proteins and genes use the same name



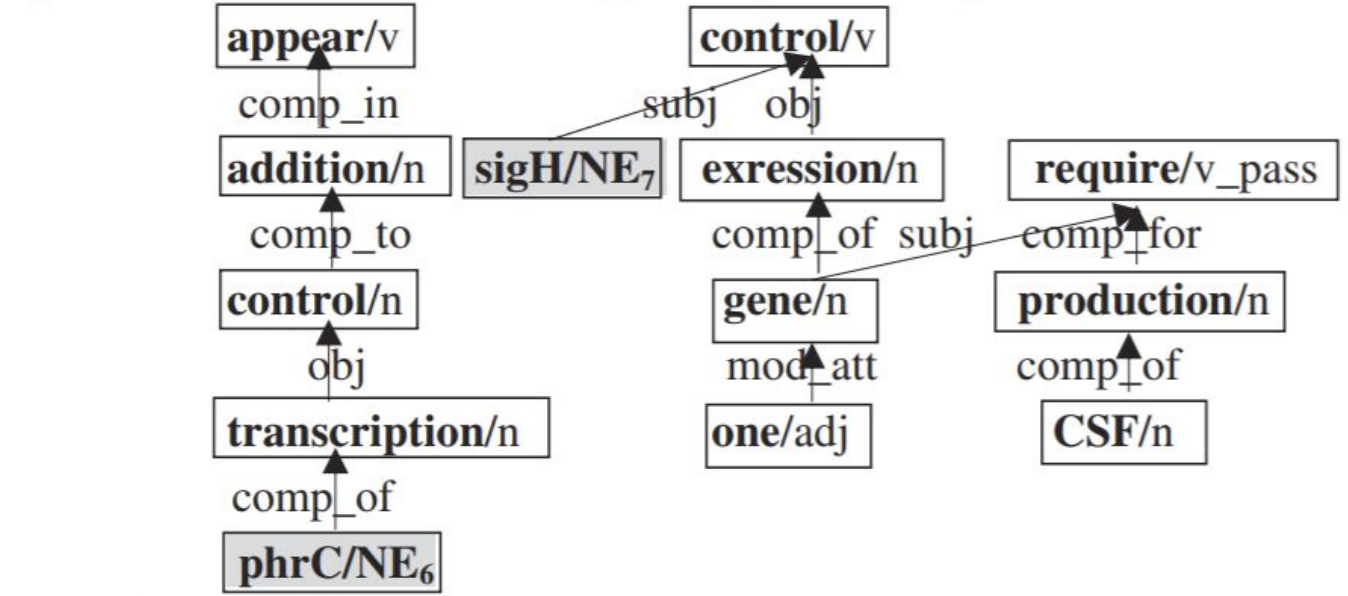


Extracting Protein-Protein/Gene-Gene Interactions

- Efforts have been made to extract **semantic relationships** between **proteins automatically from free text on PubMed** [Kim et al., 2007]
- Similar analysis methods of literature abstracts include **grammatical analysis** [Fundel et al., 2006, Šaric et al., 2005], **context-free grammar analysis** [Temkin and Gilder, 2003], and other information retrieval methods.
- Recently, another method was proposed in [Mallory et al., 2015] called **DeepDive** for large scale extraction of gene interaction from full-text literature

Extracting Protein-Protein/Gene-Gene Interactions

b In addition to controlling transcription of phrC, sigmaH appears to control expression of at least one other gene required for production of CSF.



genic_interaction(7,6)

Extracting Gene-Trait Interactions (Predicting Genes)

- Various different approaches have been proposed for this.
  - SUSPECTS [Adie et al., 2006] is a web-based server which combines annotation and sequence-based approaches to prioritize disease candidate genes in large regions of interest.
  - POCUS [Turner et al., 2003] (prioritization of candidate genes using statistics), a novel computational approach to prioritize candidate disease genes
- Recently, another method was proposed in [Gaulton et al., 2007] called CAESAR, which tries to address some of the challenges with SUSPECTS and POCUS.
- Particular, CAESAR is not limited to one or more genomic regions, and all genes annotated in database are candidates.

SUSPECTS
CANDIDATE GENE SEARCH

home search help v28.3

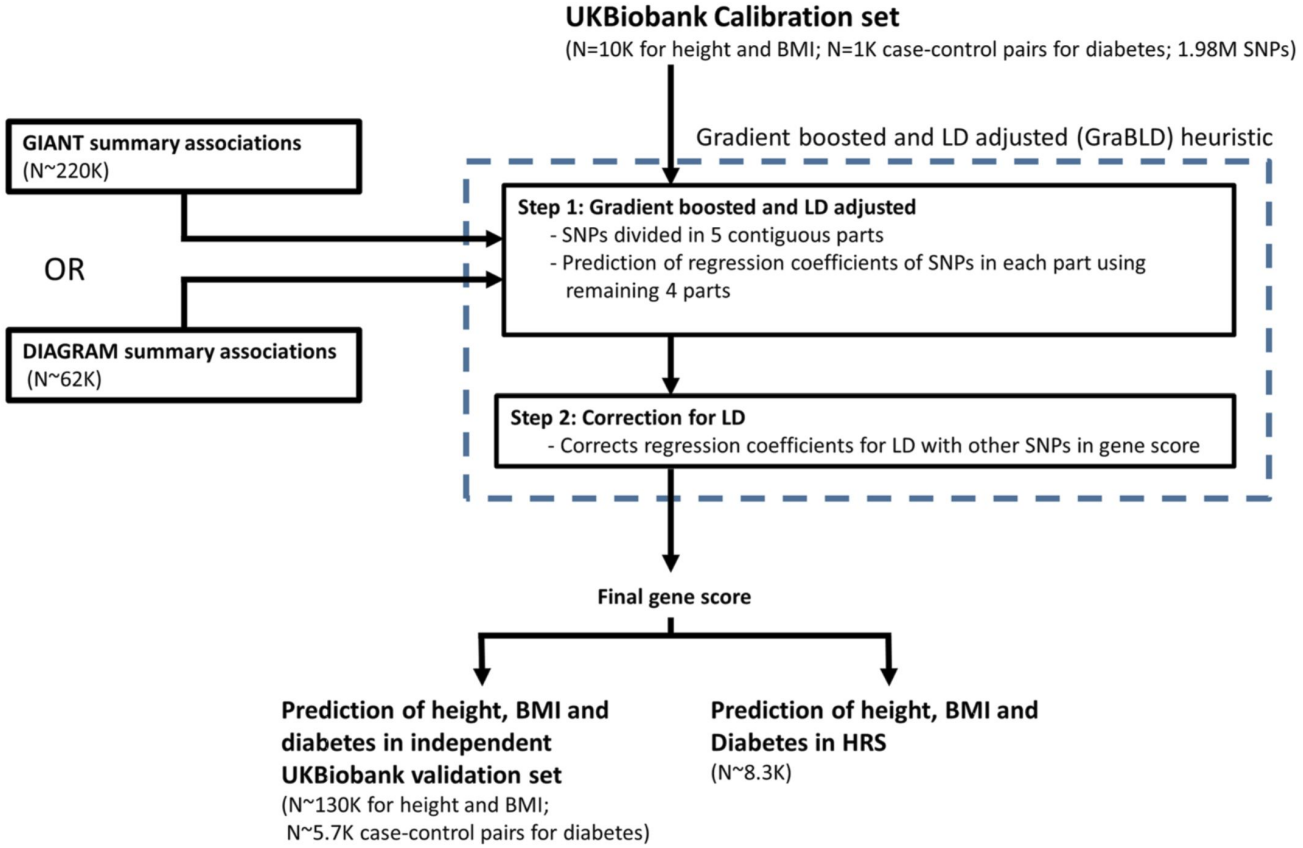
Rank genes between markers | Rank genes around a marker | Rank genes fulfilling criteria

- **obesity QTL** (see reference)
- **rheumatoid arthritis QTL** (see reference)
- **asthma QTL** (see reference)
- **deafness QTL** (see reference)
- **premature degenerative osteoarthropathy (Mendelian)** (see reference),

Extracting Trait-Gene Interactions (Predicting Traits)

- The relationship can also be **extracted the other way round.**
- Given a summary statistic of genes and polygenic traits as training examples, **can a system be trained to produce polygenic traits given a gene profile**?
- Recently Paré et al. in [Paré et al., 2017] proposed a **machine learning based method** which uses summary-level data from large genome-wide meta-analyses to improve the prediction of polygenic traits.

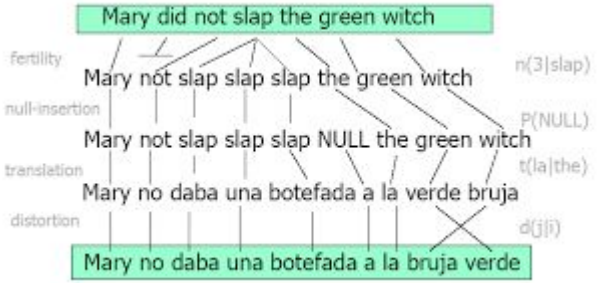
Extracting Trait-Gene Interactions (Predicting Traits)



Applications of Text Mining in DNA/RNA Sequence Prediction

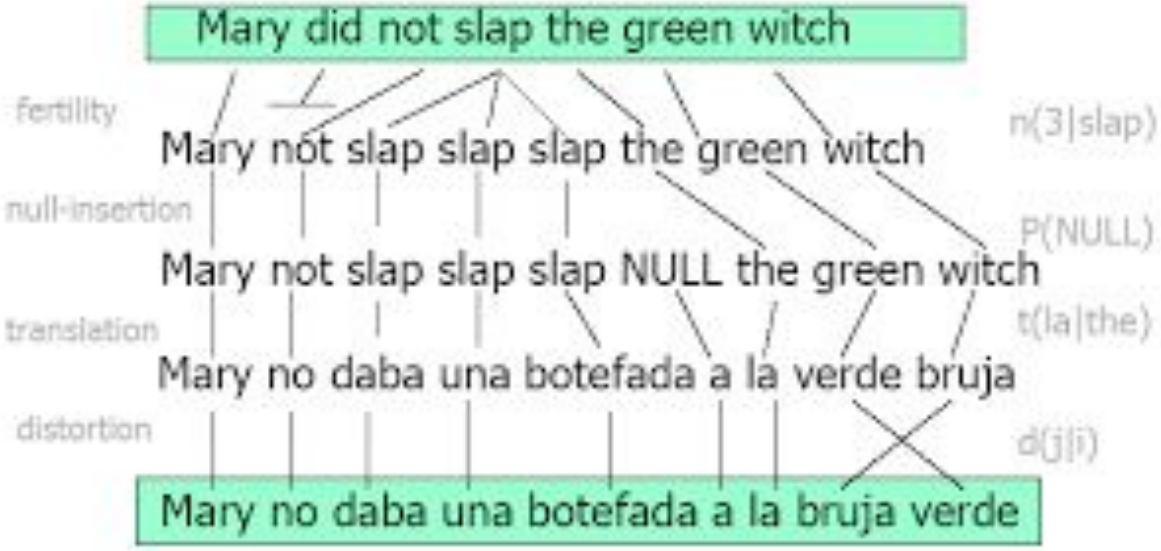
Sequence Alignment in DNA/RNA

- Alignment is also an important topic in natural language processing.
- DNA or RNA sequences can also be viewed as text. Sequence-based multiple sequence alignment methods can be used only at the sequence similarity level.
- The secondary structures of ncRNAs are usually more conserved than their sequences
- The functions of many ncRNAs are therefore determined by their secondary structure rather than by their sequences.
- As a result, structure-based multiple sequence alignment methods have been developed to align an input sequence to known ncRNA structures to determine the ncRNA class to which the input sequence belongs



Sequence Alignment in DNA/RNA

- Alignment is also an important topic in natural language processing.
- It was used for an early version of machine translation system.



Sequence Alignment in DNA/RNA

- To this end, LocARNA [Will et al., 2007] have been proposed and used to automatically align secondary structures using an energy function.
- Another extension to this method called LocRNA-P [Will et al., 2012] uses a probabilistic model to model uncertainities in sequence alignment.

Remarks and Conclusion

Remarks

- Many methods proposed for NLP tasks such as NER, POS Tagging, Semantic Relationship mapping and so on can be directly applied to a number of biomedical text analysis problems.
- Some of these methods can be generalized to work with DNA sequences, performing sequence alignment, motif detection, and protein structure prediction.

Future Work

- Where are we headed?
  - NLP in biomedical text analysis have enabled industry to spur into action.
    - Automated report generation, where layman language from patient-doctor interaction is automatically mapped to proper biomedical concepts in UMLS dictionary.
      - Echo -> echocardiogram
      - Valve in aorta -> Aortic Valve Stenosis
    - Many startups are working in this domain
      - Phraze https://www.phraze.co/
      - PlutoHI http://plutohi.com
      - NotableHealth http://notablehealth.com/
  - Easy and highly relevant searchability from an ocean of free text literature
    - Google like capability to search through unstructured free text, while producing results similar to structured database queries.
    - Exact results for text-based queries.
    - Query: “Does p53 pathway get suppressed in high grade gliomas?”
      - Answer: “Yes. More details can be found in paper: Modern Brain Tumor Imaging”

Future Work

- Inspiration from other fields
  - As more and more NLP methods are being improved using Machine Learning (ML) and Deep Learning (DL), it is only a matter of time before ML and DL enters into bioinformatics completely.
  - Even for tasks not directly related to NLP, ML/DL can prove useful.
  - As more gold-standard data is generated (DNA sequence alignment, gene-trait relationships), data-driven technologies like ML/DL will start becoming relevant.
  - However, ML/DL applications may pose a challenge regarding interpretability.

Some Applications not Covered in this Review

Some Applications Not Covered in this Review

- Text Mining applications in Protein Research
- Application to noncoding RNA Identification
- A good place to review these applications is [Zeng et al., 2015]

References

References

Adie, E. A., Adams, R. R., Evans, K. L., Porteous, D. J., and Pickard, B. S. (2006).
SUSPECTS : enabling fast and effective prioritization of positional candidates.
Bioinformatics, 22(6) :773–774.

Altman, R. B., Bergman, C. M., Blake, J., Blaschke, C., Cohen, A., Gannon, F., Grivell, L., Hahn, U., Hersh, W., Hirschman, L., Jensen, L., Krallinger, M., Mons, B., O'Donoghue, S. I., Peitsch, M. C., Rebholz-Schuhmann, D., Shatkay, H., and Valencia, A. (2008).
Text mining for biology - the way forward : opinions from leading scientists.
Genome Biology, 9(Suppl 2) :S7.

Bader, G. D., Betel, D., and Hogue, C. W. V. (2003).
BIND : the biomolecular interaction network database.
Nucleic acids research, 31(1) :248–250.

Bodenreider, O. (2004).
The unified medical language system (UMLS) : integrating biomedical terminology.
Nucleic acids research, 32(suppl_1) :D267—-D270.

References

Bosco, G. L. and Di Gangi, M. A. (2016).
Deep learning architectures for DNA sequence classification.
In *International Workshop on Fuzzy Logic and Applications*, pages 162–171. Springer.

Chou, P. Y. and Fasman, G. D. (1978).
Empirical predictions of protein conformation.
Annual review of biochemistry, 47(1) :251–276.

Dang, T. H., Le, H.-Q., Nguyen, T. M., and Vu, S. T. (2018).
D3NER : Biomedical named entity recognition using CRF-biLSTM improved with fine-tuned embeddings of various linguistic information.
Bioinformatics, 1 :8.

Du, J., Rozowsky, J. S., Korbel, J. O., Zhang, Z. D., Royce, T. E., Schultz, M. H., Snyder, M., and Gerstein, M. (2006).
A supervised hidden markov model framework for efficiently segmenting tiling array data in transcriptional and chIP-chip experiments : systematically incorporating validated biological knowledge.
Bioinformatics, 22(24) :3016–3024.

References


Fukuda, K.-i., Tsunoda, T., Tamura, A., Takagi, T., and Others (1998).
Toward information extraction : identifying protein names from biological papers.
In *Pac symp biocomput*, volume 707, pages 707–718.


Fundel, K., Küffner, R., and Zimmer, R. (2006).
RelEx—Relation extraction using dependency parse trees.
Bioinformatics, 23(3) :365–371.


Gaulton, K. J., Mohlke, K. L., and Vision, T. J. (2007a).
A computational system to select candidate genes for complex human traits.
Bioinformatics, 23(9) :1132–1140.


Gaulton, K. J., Mohlke, K. L., and Vision, T. J. (2007b).
A computational system to select candidate genes for complex human traits.
Bioinformatics, 23(9) :1132–1140.


Hatzivassiloglou, V., Duboue, P. A., and Rzhetsky, A. (2001).
Disambiguating proteins, genes, and RNA in text : a machine learning approach.
Bioinformatics, 17(suppl_1) :S97—-S106.

References

Hermjakob, H., Montecchi-Palazzi, L., Lewington, C., Mudali, S., Kerrien, S., Orchard, S., Vingron, M., Roechert, B., Roepstorff, P., Valencia, A., and Others (2004).
IntAct : an open source molecular interaction database.
Nucleic acids research, 32(suppl_1) :D452—-D455.

Kazama, J., Makino, T., Ohta, Y., and Tsujii, J. (2002).
Tuning support vector machines for biomedical named entity recognition.
In *Proceedings of the ACL-02 workshop on Natural language processing in the biomedical domain-Volume 3*, pages 1–8. Association for Computational Linguistics.

Kim, S., Yoon, J., and Yang, J. (2007).
Kernel approaches for genic interaction extraction.
Bioinformatics, 24(1) :118–126.

Krallinger, M., Leitner, F., and Valencia, A. (2010).
Analysis of Biological Processes and Diseases Using Text Mining Approaches.
pages 341–382. Humana Press.

References


Krallinger Alfonso Valencia, M., Alonso-Allende Erhardt Bioalma, R. S., Krallinger, M., Alonso-Allende Erhardt, R., and Valencia, A. (2005).
Information resources for text mining Text-mining applications integrate a broad spectrum of heterogeneous data resources, providing tools for the analysis, extraction and visualization of Text-mining approaches in molecular biology and biomedicine.
Technical Report 6.


Krauthammer, M., Rzhetsky, A., Morozov, P., and Friedman, C. (2000).
Using BLAST for identifying gene and protein names in journal articles.
Gene, 259(1) :245–252.


Kwon, D., Kim, S., Shin, S.-Y., Chatr-aryamontri, A., and Wilbur, W. J. (2014).
Assisting manual literature curation for protein-protein interactions using BioQRator.
Database : the journal of biological databases and curation, 2014.


Le, Q. and Mikolov, T. (2014).
Distributed representations of sentences and documents.
In *International Conference on Machine Learning*, pages 1188–1196.

References

Mallory, E. K., Zhang, C., Ré, C., and Altman, R. B. (2015a).
Large-scale extraction of gene interactions from full-text literature using DeepDive.
Bioinformatics, 32(1) :106–113.

Mallory, E. K., Zhang, C., Ré, C., and Altman, R. B. (2015b).
Large-scale extraction of gene interactions from full-text literature using DeepDive.
Bioinformatics, 32(1) :btv476.

McDonald, R. T., Winters, R. S., Mandel, M., Jin, Y., White, P. S., and Pereira, F. (2004).
An entity tagger for recognizing acquired genomic variations in cancer literature.
Bioinformatics, 20(17) :3249–3251.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013).
Efficient estimation of word representations in vector space.
arXiv preprint arXiv :1301.3781.

Navarro, G. (2001).
A guided tour to approximate string matching.
ACM computing surveys (CSUR), 33(1) :31–88.

References


Paré, G., Mao, S., and Deng, W. Q. (2017).
A machine-learning heuristic to improve gene score prediction of polygenic traits.
Scientific reports, 7(1) :12665.


Šarić, J., Jensen, L. J., Ouzounova, R., Rojas, I., and Bork, P. (2005).
Extraction of regulatory gene/protein networks from Medline.


Tanabe, L. and Wilbur, W. J. (2002).
Tagging gene and protein names in biomedical text.
Bioinformatics, 18(8) :1124–1132.


Temkin, J. M. and Gilder, M. R. (2003).
Extraction of protein interaction information from unstructured text using a context-free grammar.
Bioinformatics, 19(16) :2046–2053.


Turner, F. S., Clutterbuck, D. R., and Semple, C. A. M. (2003).
POCUS : mining genomic sequence annotation to predict disease genes.
Genome biology, 4(11) :R75.

References

Wang, X., Zhang, Y., Ren, X., Zhang, Y., Zitnik, M., Shang, J., Langlotz, C., and Han, J. (2018a).
Cross-type Biomedical Named Entity Recognition with Deep Multi-Task Learning.
arXiv preprint arXiv :1801.09851.

Wang, Y., Liu, S., Afzal, N., Rastegar-Mojarad, M., Wang, L., Shen, F., Kingsbury, P., and Liu, H. (2018b).
A comparison of word embeddings for the biomedical natural language processing.
Journal of biomedical informatics.

Wei, C.-H., Kao, H.-Y., and Lu, Z. (2013).
PubTator : a web-based text mining tool for assisting biocuration.
Nucleic acids research, 41(Web Server issue) :W518–22.

Wheeler, D. L., Church, D. M., Federhen, S., Lash, A. E., Madden, T. L., Pontius, J. U., Schuler, G. D., Schriml, L. M., Sequeira, E., Tatusova, T. A., and Others (2003).
Database resources of the National Center for Biotechnology.
Nucleic acids research, 31(1) :28–33.

References


Wilbur, W. J. and Coffee, L. (1994).
The effectiveness of document neighboring in search enhancement.
Information Processing & Management, 30(2) :253–266.


Will, S., Joshi, T., Hofacker, I. L., Stadler, P. F., and Backofen, R. (2012).
LocARNA-P : accurate boundary prediction and improved detection of structural RNAs.
Rna.


Will, S., Reiche, K., Hofacker, I. L., Stadler, P. F., and Backofen, R. (2007).
Inferring noncoding RNA families and classes by means of genome-scale structure-based clustering.
PLoS computational biology, 3(4) :e65.


Xenarios, I., Salwinski, L., Duan, X. J., Higney, P., Kim, S.-M., and Eisenberg, D. (2002).
DIP, the Database of Interacting Proteins : a research tool for studying cellular networks of protein interactions.
Nucleic acids research, 30(1) :303–305.


Yip, K. Y., Cheng, C., and Gerstein, M. (2013).
Machine learning and genome annotation : a match meant to be ?
Genome biology, 14(5) :205.

References

Zeng, Z., Shi, H., Wu, Y., and Hong, Z. (2015).
Survey of natural language processing techniques in bioinformatics.
Computational and mathematical methods in medicine, 2015.