

ChemTextMiner: An open source tool kit for mining medical literature abstracts

M.Karthikeyan^{*1}, Yogesh Pandit¹, Deepak Pandit¹, Ganesh Nainaru¹, Sunil Nalwade¹, Renu Vyas², Esha Jain²

¹DIRC, National Chemical Laboratory, Pune- 411008, ² Dr. D.Y. Patil Biotechnology and Bioinformatics Institute Pune- 4110033

karthincl@gmail.com_976427981



ABSTRACT: Text mining involves recognizing pattern from a wealth of information hidden latent in unstructured text and deducing explicit relationship among data entities by using data mining tools. Text mining of Biomedical literature is essential for building biological network connecting genes, proteins, drugs, therapeutic categories, side effects etc. related to diseases of interest. We present an approach for textmining biomedical literature mostly in terms of not so obvious hidden relationships and build biological network and was applied for the textmining of important human diseases like MTB, Malaria, Alzheimer and Diabetes. The methods, tools and data used for building biological network using distributed computing environment previously used for ChemXtreme[1] and ChemStar[2] applications are also described.

MATERIALS: ChemTextMiner was completely build on Java platform, LingPipe, Genia Corpus for medical data mining, MySQL, RapidMiner (machine learning tool for data classification) CytoScape for integrating, visualizing, data in the form of networks, JPROLINE for distributed computing.

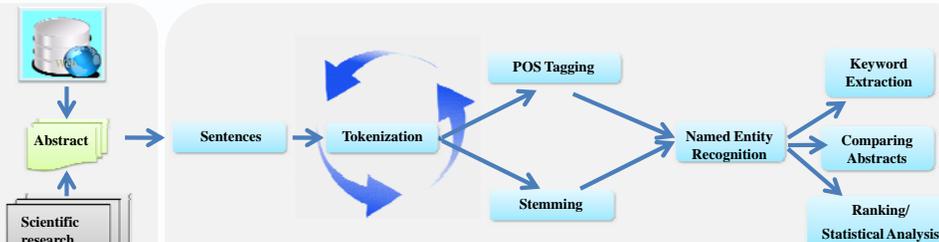


Fig: 1 Flowchart showing the entire methodology implemented by ChemTextMiner.

Lingpipe recognizes these set of words as belonging to class 'Protein molecule'

ChemTextMiner is better than lingpipe due to its inbuilt capability to recognize biological and chemical semantics as is evident from above figure

RESULTS AND DISCUSSION:

- ✓ The data in each class was relevant to the disease and showed less ambiguity.
- ✓ The case studies were done on different diseases like Diabetes, Alzheimer's and MTB where we found appropriate results.
- ✓ The abstracts were stored in the local database and passed through the ChemTextMiner to find the disease related proteins, organic and inorganic molecules.
- ✓ The results were shown below in tabular and network formats.

Table 1: Top ranked entries from the nine protein classes.

PROTEIN CLASSES	COUNT	PROTEIN CLASSES	COUNT
Amino_acid_monomer	25226	Protein_family	186491
Peptide	53260	Protein_molecule	457418
Protein_complex	34753	Protein_N	9016
Protein_domain	24448	Protein_substructure	4000
Protein_subunit	7205		

Table 2: Calculated Network properties using Cytoscape

Parameters	Value
Network Nodes	121
Network Heterogeneity	2.591
Network Density	0.017
Network Diameter	6
Network centralization	0.246
Avg. number of neighbors	1.983

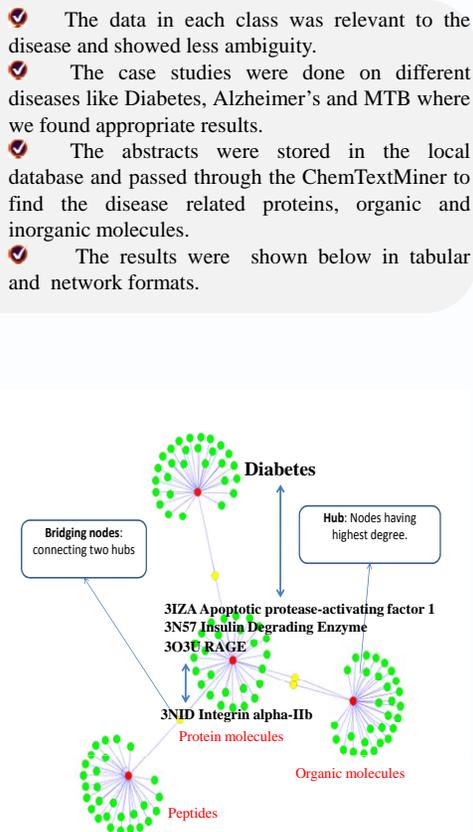


Fig 3: Interaction Network revealing not so obvious hidden relationships

CONCLUSION:

We have produced a comprehensive, fast, and extensible tool ChemTextMiner for extracting Biological information from massive data sets and identification of unknown relationships between the extracted subjects. The ChemTextMiner is helpful in multiple research problems like protein-protein interaction studies, drug discovery and chemical library creation etc. The case study on the Diabetes disease suggested that the data extracted by the tool was showing less ambiguity and more promiscuity. The network analysis on resultant data revealed some hidden relations between the classes which may be useful in solving some of biological problems which are not obvious without high-throughput text mining methodologies.

ACKNOWLEDGEMENT:

MK thanks CSIR for financial support. RV thanks DYPBBI for infrastructural support. The authors declare no conflict of interest.

REFERENCES:

- [1]Harvesting Chemical Information from the Internet Using a Distributed Approach:ChemXtreme (2006) J. Chem. Inf. Model., 46 (2), 452-46 1.
- [2] Distributed Chemical Computing Using ChemStar: Open Source Java RMI Architecture applied to Large Scale Molecular Data from PubChem. (2008) J. Chem. Inf. Model., 48 (4), 691-703.