Mining Disease Associated Biomarker Networks from PubMed

Zhong Huang
School of Information Science and Technology
Drexel University
Philadelphia, USA

Abstract—**Disease related biomarker discovery is the critical step to realize the future personalized medicine and has been an important research area. With exponential growing of biomedical knowledge deposited in PubMed database, it is now an essential step to mine PubMed for biomarker-disease associations to support the laboratory research and clinical validation. We constructed list of human diseases that are most frequently associated with biomarker in literatures by text mining. Top ranked neurology diseases were then used to extract associated genes from PubMed using context sensitive information retrieval methods. Associated genes were then integrated into pathways and subject to network biomarker analysis. Our approach identifies both known and potential biomarkers for 3 neurodegenerative diseases.**

Keywords—biomarker; disease-gene association; semantic; text mining; biological network

I. INTRODUCTION

During the past decades, high-throughput proteomics techniques have been widely employed for identifying disease associated genes, proteins, and metabolites. It led to rapidly accumulated experiment data and research reports. Identifying biomarkers and their interaction network underlying different diseases has become an important step to realize the future personal medicine. Based on NIH definition the biomarker is a wide range of markers that can be objectively measured and evaluated to indicate normal biological or pathogenic processes [1]. To aid biomarker discovery the text mining technique has been utilized to analyze heterogeneous data sources. PubMed database comprises 21 million literature citations in biomedical fields and have been undergoing rapid update with growing experimental data analysis from high-throughput -omics study. To develop an efficient text mining approach to reveal underlying disease associated biomarkers from huge amount of literature are therefore extremely needed. Biomarkers show significant diversity ranging from genes, proteins, nucleic acid, and small metabolites, and have been applied throughout disease prediction, prognosis, and during various stages of drug discovery. Moreover, due to the nature of high variability of gene, protein, and disease used in biomedicine literature reports, semantic search and information retrieval played an important role in biomarker discovery. Named Entity Recognition (NER) combined with semantic annotation of biological entities including domain specific ontology, dictionary and thesaurus are often used to extract biological entities from text in order to achieve high accuracy and recall. Biomarker candidates discovery is considered as discovery of hidden semantic relations between diseases and genes. List of candidate biomarkers mined from literature can then be subject to further empirical validation before a novel biomarker can be applied.

With the advancement of text mining technology and rapid accumulation of proteomics data, more and more researches have been focused on finding potential biomarker candidates from literature database as the first step of biomarker discovery. Web based tools including PolySearch [2], iHop [3], EBIMed [4], and Semedico [5] are four representative systems focusing on biological entity associations mining from biomedical literatures. However above methods are largely based on dictionary approach which doesn't take into account the contextual and semantic relations between potentially associated biological entities. In this paper, we extracted diseases associated with biomarker discovery from PubMed to reveal biomarker discovery activity in different areas. We then applied semantic based information retrieval technique to mine PubMed for genes associated with specific diseases.

II. MATERIALS AND METHODS

A. *Construction list of neurology diseases associated with biomarkers from PubMed*

List of human diseases associated with biomarker was extracted from PubMed using PolySearch [2]. The PolySearch algorithms query PubMed and by parsing the retrieved abstracts it extracts list of scored diseases associated with the query term using 'bag-of-words' approach. This approach depends on a comprehensively compiled human disease and gene/protein thesaurus containing 25944 normalized names and synonyms derived from UMLS, OMIS, and manual curation. Top ranked biomarker associated human diseases related to neurology disease with z-score cutoff 0.1 were selected for downstream analysis described below. The list of chosen diseases are regarded as high degree of biomarker research topics in literature report. Among those highly cited

978-1-4799-1389-3/13/$31.00 ©2013 IEEE

neurology diseases, Alzheimer disease, multiple sclerosis, and Parkinson's disease are selected for further text mining analysis in an aim to retrieve their associated genes as disease biomarker candidates.

B. *Extract specific disease associated genes from PubMed*

To extract specific disease associated genes from PubMed database we implemented Java program using Dragon Tool kit [6], an information retrieval and text mining framework. The toolkit provides built-in semantic based language models for information retrieval and text mining. We also implemented a Windows user interface using C# to wrap around the toolkit Java package using IKVM Java Virtual Machine to supply an user-friendly interface for querying and indexing PubMed. Currently the system implemented 3 modules for Dragon toolkit xml configuration, online literature indexing, and information retrieval.

PubMed abstracts related to Alzheimer disease, multiple sclerosis, and Parkinson's disease were retrieved from PubMed dynamically using online collection writer of Dragon toolkit. The document collection include all publications with title and abstract, returned from PubMed API by semantic query of disease Medical Subject Headings (MeSH) descriptors. MeSH terms were manually assigned to each paper when it is deposited to the PubMed to represent its full-text topic. The document collection for Alzheimer disease, Multiple sclerosis, and Parkinson disease contains 49943, 28419, and 31321 abstracts respectively.

In this work, we used a context-sensitive semantic language modeling approach to identify biological associations within natural text [7]. This approach allows explicit extraction of topic signature as concept pair from document using ontology concepts, and map them onto single-word features based on co-occurrence data for improved information retrieval. The topic signature pair consists two concepts that are semantically and syntactically related to each other thus significantly disambiguated biological terms in the text [8]. Two ontology thesauri, namely UMLS and MeSH, are originally implemented in the toolkit. To adopt the toolkit in disease gene association mining, due to limitations of UMLS and MeSH in gene and protein named entity recognition, we incorporated GO ontology for concept lookup. The GO ontology defines controlled vocabulary for gene related terms along with relationships between them [9] and has been widely applied in various of biomedical data mining tasks. After applying stop word list, part of speech (POS) tagging, the document is then extracted for gene concepts based on GO ontology thesaurus. Documents were indexed using concept based indexing and the sparse matrix was created subsequently.

C. *Disease - gene network construction*

List of genes associated with specific neurology disease are ranked by their co-occurrence with each other and assigned the unique Entrez gene Id through Entrez eUtility API. The disease-associated gene network was build using atBioNet [10]. The tool constructs the interact network by modified SCAN algorithm enabling enrichment analysis and assessment of generated gene network for discovery of biomarker candidates.

III. Results and Discussion

A. *Diseases associated with biomarkers*

Total 347 diseases that are associated with biomarker were retrieved from PubMed using PolySearch with query pattern "given biomarker find all associated diseases". Retrieved diseases are ranked by relevance score representing all diseases found with biomarker associations from PubMed. We then classified the top 100 scored diseases using high-level Disease Ontology [11]. Figure 1 shows the classification of diseases associated with biomarkers mined from PubMed using disease ontology.

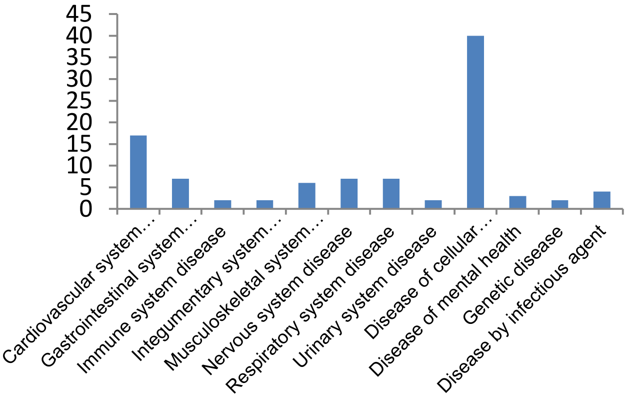

Figure 1. Disease Ontology classification of 347 biomarker associated human diseases mined from PubMed.

As shown in figure 1, it is evident that biomarker discovery associated with cancer, e.g. disease of cellular proliferation, are the most intensively studied areas, followed by cardiovascular system disease, gastrointestinal system disease, and nervous system disease.

B. *Retrieval of gene disease associations*

To further explore potential biomarkers associated with diseases that have most citations in biomarker research, we selected 3 neurology diseases from 347 diseases, and extracted associated genes from PubMed by semantic based information retrieval approach. Table 1 shows the 3 neurodegenerative diseases and their disease ontology and MeSH id.

Name	DOID	Mesh ID
Alzheimer's disease	10552	MSHD000544
Multiple sclerosis	2377	MSHD009103
Parkinson's disease	14330	MSHD010300

Table 1. List of neurology diseases gene associated with biomarker discovery mined from PubMed. The Disease

Alzheimer's disease		Multiple sclerosis		Parkinson's disease	
Gene	Entrez Id	Gene	Entrez Id	Gene	Entrez Id
APP	351	MOG	4340	LRRK2	120892
PSEN1	5663	IFNB1	3456	PINK1	65018
BACE1	23621	LILRA3	11026	SNCAIP	9627
PSEN2	5664	MOBP	4336	SNCA	6622
STH	246744	MBP	4155	PARK7	11315
MAPT	4137	PLP1	5354	FGF20	26281
YWHAQ	10971	CLDN11	5010	SNCB	6620
NCSTN	23385	GPC5	2262	PARK2	5071
PHF1	5252	AQP4	361	GBA	2629
APBB1	322	LINGO1	84894	NR4A2	4929
APLP2	334	COP1	114769	ATXN2	6311
APLP1	333	PADI2	11240	UBC	7316
COL25A1	84570	IL26	55801	MCF2L	23263
ITM2B	9445	TALDO1	6888	SLC6A3	6531
APBB3	10307	MX1	4599	UCHL1	7345
APOE	348	PADI4	23569	APOE	348
CTNNA3	29119	HAVCR2	84868	TARDBP	23435
APH1B	83464	MAG	27307	GDNF	2668
DLST	1743	C4orf6	10141	COMT	1312
TM2D1	83941	IL23R	149233	SLC18A2	6571

Ontology ID (DOID) and Mesh Id are assigned for each disease.

Genes associated with Alzheimer's disease, Multiple sclerosis, and Parkinson's disease were retrieved and mapped to Entrez gene Id. This yielded a list of 297, 185, and 49 genes associated with Alzheimer's disease, Multiple sclerosis, and Parkinson's disease respectively. Table 2 shows the top 20 disease associated genes and their corresponding Entrez gene Id.

Table 2. List of 20 genes associated with Alzheimer's disease, Multiple sclerosis, and Parkinson's disease mined from PubMed literature database.

Alzheimer's disease is a neurodegenerative disease accounting for 60% of all dementia diagnosed clinically [12]. Biomarkers from plasma, urine, CSF have been reported as Alzheimer's disease diagnosis and indication of disease progression. Among those well established biomarkers that are validated in laboratory or clinical study, APP, BACE1, MAPT, α-synuclein, TARDBP have been retrieved by our text mining approach [13].

Multiple sclerosis is an autoimmune disease with hallmark of demyelination in brain and spinal cord. Polymorphisms in Major Histocompatibility Complex HLA class II antigens have been regarded as the main genetic risk factor for multiple sclerosis. In agreement with those reports, we found HLADRB1, HLADRB4, HLADRB5, and HLADQB1 were retrieved as multiple sclerosis biomarker candidates. Expression of myelination related genes including MOG, MBP, PLP1, MAG have been shown abnormal at the onset and during disease progress [14]. However reports about antibodies against myelin related genes are controversial thus their diagnostic and prognostic values remains to be confirmed.

Parkinson's disease is a neurological degenerative disorder caused by destruction of dopamine-generating cells in the midbrain. Our results show several genes mined from PubMed were reported as biomarker candidates in Parkinson's disease. SNCA is one of the most studied and validated biomarker that plays critical role in pathogenesis of synucleinopathies and neurodegeneration [15].

C. *Construct the disease gene network*

To get insight into potential biomarker candidates that are involved in pathogenesis of disease, we connected the multiple sclerosis associated genes using atBioNet. We found 172 out of 185 input genes were successfully mapped to 6 major modules using the stringent node addition option which adding only nodes directly connected to at least two input nodes. Total 642 genes from KEGG human database were added to network and connected by identified pathways.

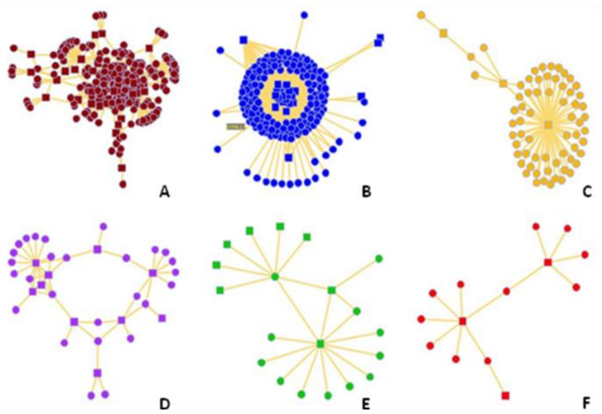

Figure 2. Interact network constructed from the multiple sclerosis associated genes. The seed genes are square shaped and added genes are circle points. Major modules identified are shown in A-F.

Pathway analysis shows module A contains ErbB, cytokine, chemokin, and apoptosis related signaling pathways. Module B includes cytokin-cytokin receptor interaction network and glutamatergic, cholinergic synapse pathways. Module C and D contain mainly inflammatory pathways. Nervous system and neurodegenerative pathways were identified in module E. GFAP and MOG found in module F have been reported as prognostic biomarkers.

IV. CONCLUSION

In this paper we introduced a context sensitive semantic based information retrieval method to mine disease associated gene biomarker candidates from literature database, as the first step towards developing an retrieval and in-silico validation system for disease associated biomarker candidates. Our experiment results demonstrated the effectiveness of our semantic based contextual sensitive indexing and information retrieval. Using this approach we are able to construct an informative gene network of neurodegenerative diseases through text mining of PubMed abstracts.

REFERENCES

[1] Biomarkers Definitions Working Group. Review Biomarkers and surrogate endpoints: preferred definitions and conceptual framework. Clin Pharmacol Ther. 2001 Mar; 69(3):89-95.

[2] Dean Cheng, Craig Knox, Nelson Young, Paul Stothard, Sambasivarao Damaraju, David S. Wishart. "PolySearch: a web-based text mining system for extracting relationships between human diseases, genes, mutations, drugs and metabolites". Nucleic Acids Res. 2008 July 1; 36(Web Server issue): W399–W405.

[3] Hoffmann R, Valencia A. "Implementing the iHOP concept for navigation of biomedical literature". Bioinformatics. 2005 Sep 1;21 Suppl 2:ii252-8.

[4] Rebholz-Schuhmann D, Kirsch H, Arregui M, Gaudan S, Rynbeek M, Stoehr P. "Protein annotation by EBIMed". Nat Biotechnol 2006, 24:902-903.

[5] Wermter J, Tomanek K, Hahn U. "High-performance gene name normalization with GeNo." Bioinformatics. 2009 Mar 15;25(6):815-21.

[6] Zhou, X., Zhang, X., and Hu, X., "Dragon Toolkit: Incorporating Auto-learned Semantic Knowledge into Large-Scale Text Retrieval and Mining," In proceedings of the 19th IEEE International Conference on Tools with Artificial Intelligence (ICTAI), October 29-31, 2007, Patras, Greece.

[7] Xiaohua Zhou, Xiaohua Hu, Xiaodan Zhang, Xia Lin, Il-Yeol Song.. "Context-sensitive semantic smoothing for the language modeling approach to genomics IR". in ACM SIGIR 2006, Aug 6-11

[8] Xiaohua Zhou, Xiaodan Zhang, and Xiaohua Hu, "MaxMatcher: Biological Concept Extraction Using Approximate Dictionary Lookup". in Q. Yang and G. Webb (Eds.): PRICAI 2006, LNAI 4099, pp. 1145–1149, 2006. Springer-Verlag Berlin Heidelberg 2006

[9] The Gene Ontology Consortium (January 2008). "The Gene Ontology project in 2008". Nucleic Acids Res. 36 (Database issue): D440–4.

[10] Ding Y, Chen M, Liu Z, Ding D, Ye Y, Zhang M, Kelly R, Guo L, Su Z, Harris SC, Qian F, Ge W, Fang H, Xu X, Tong W. "atBioNet--an integrated network analysis tool for genomics and biomarker discovery". BMC Genomics. 2012 Jul 20;13:325.

[11] Schriml LM, Arze C, Nadendla S, Chang YW, Mazaitis M, Felix V, Feng G, Kibbe WA. "Disease Ontology: a backbone for disease semantic integration". Nucleic Acids Res. 2012 Jan;40(Database issue):D940-6

[12] Brookmeyer R, Johnson E, Ziegler-Graham K, Arrighi HM." Forecasting the global burden of Alzheimer’s disease". Alzheimer’s and Dementia. 2007;3(3):186–191

[13] Zetterberg, H; Andreasson, U; Hansson, O; Wu, G; Sankaranarayanan, S; Andersson, ME; Buchhave, P; Londos, E et al. (2008). "Elevated cerebrospinal fluid BACE1 activity in incipient Alzheimer disease". Archives of neurology 65 (8): 1102–7.

[14] Katsavos S, Anagnostouli M. "Biomarkers in Multiple Sclerosis: An Up-to-Date Overview". Mult Scler Int. 2013;2013:340508.

[15] Borghi R, Marchese R, Negro A, Marinelli L, Forloni G, Zaccheo D, et al. "Full length alpha-synuclein is present in cerebrospinal fluid from Parkinson’s disease and normal subjects". Neurosci Lett 2000; 287: 65–7

978-1-4799-1389-3/13/$31.00 ©2013 IEEE