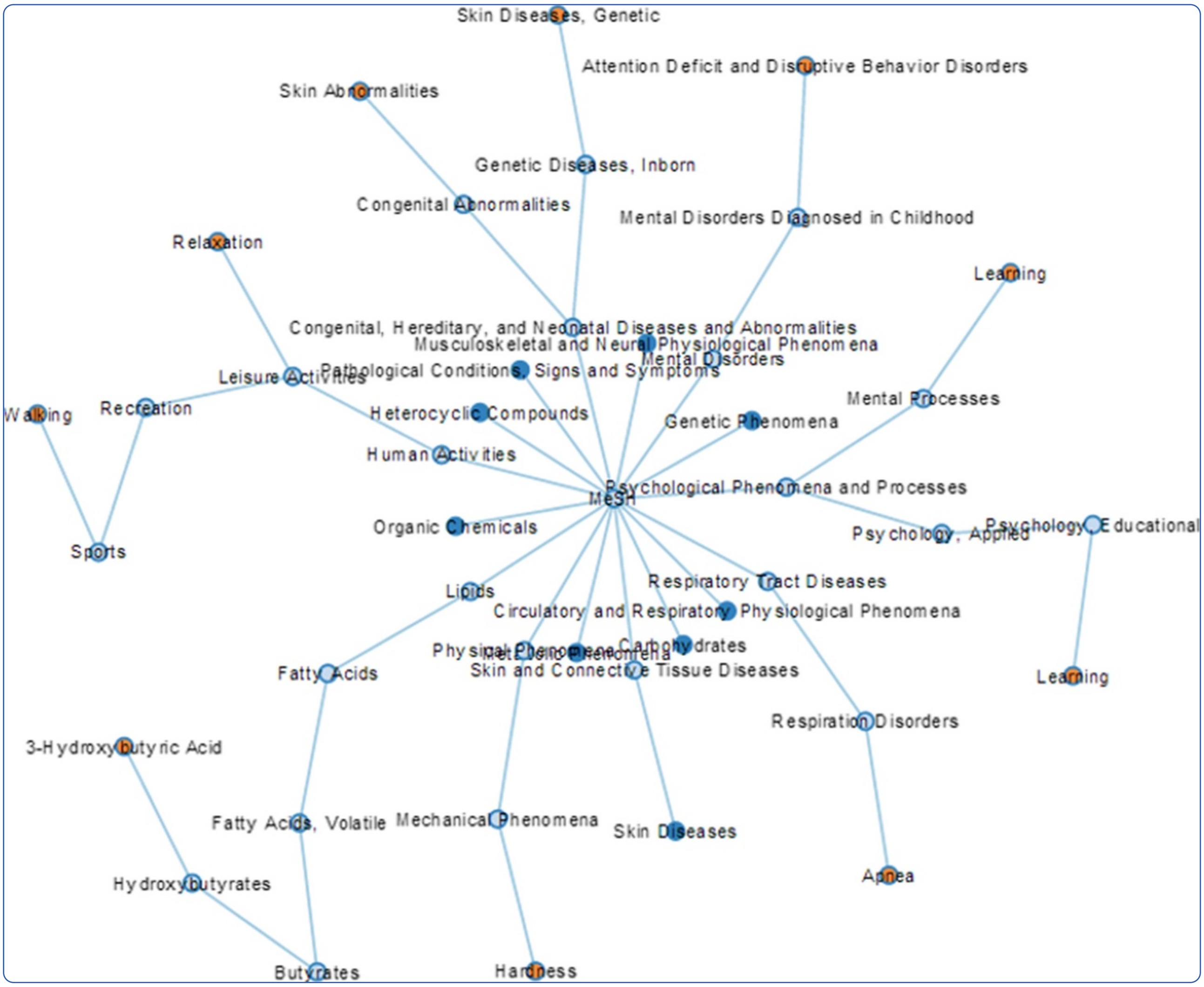
JOURNAL OF
TRANSLATIONAL MEDICINE



Empirical study using network of semantically related associations in bridging the knowledge gap

Abedi *et al.*

BioMed Central

Abedi *et al. Journal of Translational Medicine* 2014, **12**:324
http://www.translational-medicine.com/content/12/1/324

JOURNAL OF TRANSLATIONAL MEDICINE

METHODOLOGY Open Access

Empirical study using network of semantically related associations in bridging the knowledge gap

Vida Abedi[1*], Mohammed Yeasin[2,3] and Ramin Zand[4]

Abstract

Background: The data overload has created a new set of challenges in finding meaningful and relevant information with minimal cognitive effort. However designing robust and scalable knowledge discovery systems remains a challenge. Recent innovations in the (biological) literature mining tools have opened new avenues to understand the confluence of various diseases, genes, risk factors as well as biological processes in bridging the gaps between the massive amounts of scientific data and harvesting useful knowledge.

Methods: In this paper, we highlight some of the findings using a text analytics tool, called ARIANA - Adaptive Robust and Integrative Analysis for finding Novel Associations.

Results: Empirical study using ARIANA reveals knowledge discovery instances that illustrate the efficacy of such tool. For example, ARIANA can capture the connection between the drug hexamethonium and pulmonary inflammation and fibrosis that caused the tragic death of a healthy volunteer in a 2001 John Hopkins asthma study, even though the abstract of the study was not part of the semantic model.

Conclusion: An integrated system, such as ARIANA, could assist the human expert in exploratory literature search by bringing forward hidden associations, promoting data reuse and knowledge discovery as well as stimulating interdisciplinary projects by connecting information across the disciplines.

Keywords: Knowledge discovery, Hypothesis generation, Literature mining, Ontology mapping, PubMed, Medical subject headings (MeSH), Multi-gram dictionary, Latent semantic analysis (LSA), Network of association, Semantic associations

Background

Strategic reading, searching, and filtering have been the norm in gaining perspective from the "ocean of data" in the field of biomedicine. Intriguingly, the information overload has contributed in widening the knowledge gap. On an average day in 2013, approximately 3 million searches were performed on the PubMed web site, and an additional 3 million searches were done by scripts (e.g., by application programming interfaces or APIs) [1]. It is widely acknowledged that efficient mining of biological literature could provide a variety of services [2] such as assisted curation of literature [3], hypothesis generation [4], association prediction [5] or semantic sensitive knowledge discovery (Figure 1). Traditionally, literature mining tools focus on text summarization and clustering techniques [6] with the goal of reducing the data overload and ability to read and synthesize more information in a shorter time. In recent years effective LDA [4] as well as LSA [7]-based techniques have also been used with great success to formulate new hypotheses and generate connections from existing literature. It was argued that a text analytic tool capable of extracting network of semantically related associations may help in bridging knowledge gap by using human's unique visual capacity and information seeking behavior. For instance in a study, 16,169 articles were chosen to create a visual representation of main concepts, creating a visual maps of verbal

* Correspondence: abedi@vbi.vt.edu
[1]The Center for Modeling Immunity to Entering Pathogens, Nutritional Immunology and Molecular Medicine Laboratory, Virginia Bioinformatics Institute, Virginia Polytechnic Institute and State University, Blacksburg, VA 24060, USA
Full list of author information is available at the end of the article

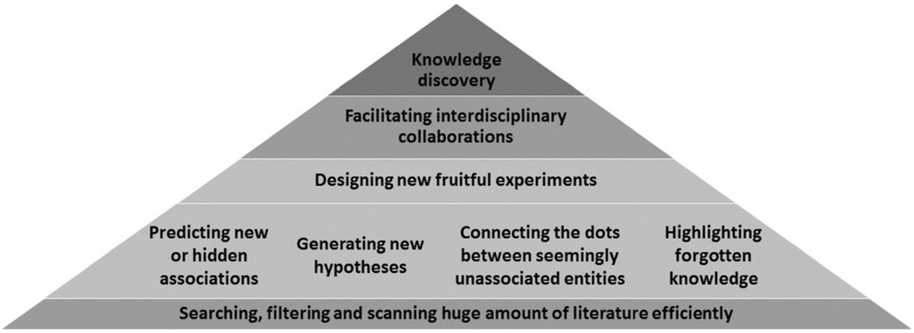

Knowledge discovery
Facilitating interdisciplinary collaborations
Designing new fruitful experiments
Predicting new or hidden associations
Generating new hypotheses
Connecting the dots between seemingly unassociated entities
Highlighting forgotten knowledge
Searching, filtering and scanning huge amount of literature efficiently

Figure 1 Array of applications of ARIANA.

information [8]. In that analysis, it was found that "*verbal presentation offers more precise information [...], whereas the visual presentation offers a more flexible style of exploration that better shows multiple, fuzzy, and intermixed and complexly patterned relations among the documents* [8]". In addition, literature mining tools that can capture the semantic relationship could in principle connect disjoints entities between different research fields.

A well designed literature mining tool can be used to further the understanding of potential confluence between various diseases, drugs, genes, proteins and other risk factors. To be effective in bridging the monotonically increasing gaps between the data production and its utilization, such a tool should efficiently map domain specific information and capture network of semantic-sensitive association and provide effective and seamless visualization (Figure 2). This paper highlights the key findings of empirical studies performed using a knowledge discovery tool, called ARIANA.

Methods

ARIANA [9] is a software system that is designed to capture *"crisp semantic associations"* among bio-medical concepts of interest and provide scalable Web-Services (Figure 2). It integrates semantic-sensitive analysis of text data through ontology mapping with database search and advanced visualization of the network of semantically related associations that can be easy to collapse and expand, allowing the user to have a global view of the results or to focus on a sub-network. As an integrative tool, goals of ARIANA are to find the network of semantic associations in bridging the gap between the production and utilization of data, disambiguate the domain specific entities, provide robust results to a broad range of queries and, deliver a scalable Web-Service using state-of-the-art technology.

Evaluating knowledge discovery tools remain challenging due to inherent subjective nature of the findings. However, empirical studies can be used to illustrate the efficacy of such tools. The semantic model of ARIANA (Figure 3) was significantly expanded to include 8,700,000

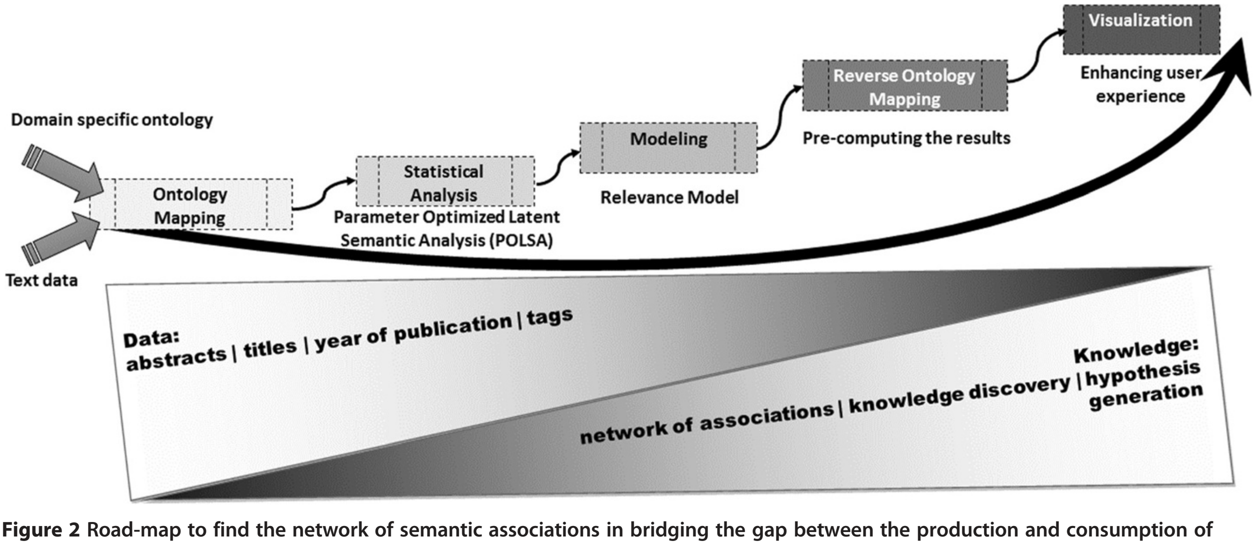

Visualization
Enhancing user experience
Reverse Ontology Mapping
Pre-computing the results
Modeling
Relevance Model
Statistical Analysis
Parameter Optimized Latent Semantic Analysis (POLSA)
Domain specific ontology
Ontology Mapping
Text data
Data:
abstracts | titles | year of publication | tags
Knowledge:
network of associations | knowledge discovery | hypothesis generation

Figure 2 Road-map to find the network of semantic associations in bridging the gap between the production and consumption of bio-medical data.

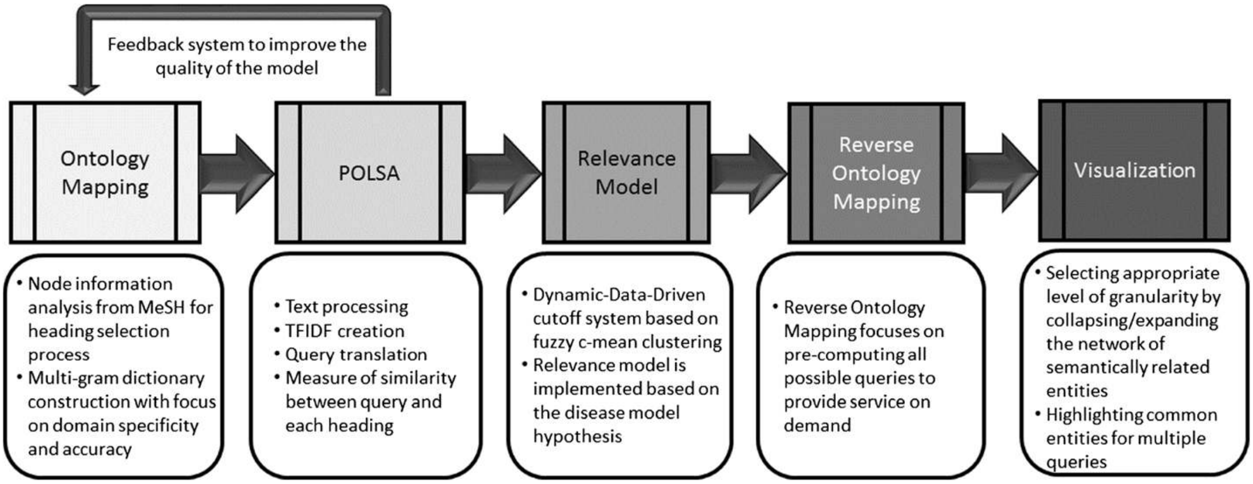

Feedback system to improve the quality of the model
Ontology Mapping
POLSA
Relevance Model
Reverse Ontology Mapping
Visualization
• Node information analysis from MeSH for heading selection process
• Multi-gram dictionary construction with focus on domain specificity and accuracy
• Text processing
• TFIDF creation
• Query translation
• Measure of similarity between query and each heading
• Dynamic-Data-Driven cutoff system based on fuzzy c-mean clustering
• Relevance model is implemented based on the disease model hypothesis
• Reverse Ontology Mapping focuses on pre-computing all possible queries to provide service on demand
• Selecting appropriate level of granularity by collapsing/expanding the network of semantically related entities
• Highlighting common entities for multiple queries

Figure 3 Data processing stages of ARIANA.

representative abstracts from the PubMed [1] database covering fifty years of literature, and 2,545 hierarchically-structured Headings[b] from MeSH (Medical Subject Headings) ontology [10]. The Heading selection process is implemented based on heuristics to capture a representative and balanced data. The main constraint being that half of the selected headings for the model have to be from the Disease category since the main focus of the study is to identify disease related risk factors. The main feature used to select the headings was the number of abstract in each category. For instance, categories with fewer than 1,000 or greater than 50,000 abstracts were not considered, as they represent very specific or very general topics. For instance, there are 1,828 headings from the Disease category (category C), 475 headings from Chemical and Drugs category (category D), and the remaining headings from categories F, G, I, J, M and N [10]. Performing Ontology Mapping (OM) on the 2012 version of the MeSH and combing to that gene symbols, from the OMIM (Online Mendelian Inheritance in Man) [11] database generated a context-specific and multi-gram dictionary with 17,074 terms (for instance, the term "yellow virus fever" is a medical multi-gram term with very specific meaning extracted from parsing the MeSH)[c]. The latter is critical for efficient mapping of domain knowledge to the semantic space with layer of genetic information. ARIANA adopts the POLSA [12] to capture direct as well as indirect statistical associations among the dictionary terms. In the POLSA model, term-frequency inverse-document-frequency (TF-IDF) matrix was used to generate the *encoding* matrix and the dimensionality was reduced to cover 95% of the total energy (dimensionality was reduced from 2,545 to 1,400 headings to create the encoding matrix). Furthermore, the multi-gram dictionary also captures higher order associations among the different entities. A Relevance Model (RM), implemented using fuzzy c-mean clustering, is introduced to logically group the ranked results of the user's query to generate hypothesis [5] and facilitate the identification of highly related entities in a Dynamic-Data-Driven (DDD) fashion [5]. Fuzzy c-means clustering is applied to group the ranked headings into three groups (highly associated, possibly associated or unknown) [9]. The cut-off values estimated through this process are DDD and subject to change as the dataset expands [9]. Reverse Ontology Mapping (ROM) was introduced to generate networks from the semantically related entities that are considered relevant. The ROM was used to map the terms back to the MeSH and create a network of associations. All possible queries were pre-computed, stored in a hash table and, linked with the visualization module in order to provide services on demand. Finally, to present the results in a graphical representation, JavaScript Object Notation (JSON*)* objects are created, and the D3 library[d] is used to implement the collapsibility and expandability as well as flexibility to stretch the objects to provide better visual clarity for each node. Figure 4 highlights the main steps in the data extraction and analysis of the system.

Results and discussion

Empirical study using the improved ARIANA was performed to identify network of associations with single as well as multiple query words. Representative of the findings are succinctly summarized below to illustrate the utility of such system in discovering unknown interactions and also to generate robust hypothesis by connecting the information from interdisciplinary fields. However, in order to extract hidden knowledge for a single vital query, such as the case for the asthma study at John Hopkins, it is imperative to not only focus on the graph representation but also extract the raw association scores and investigate entities with weaker level of associations. In essence, with no direct evidence in the literature,

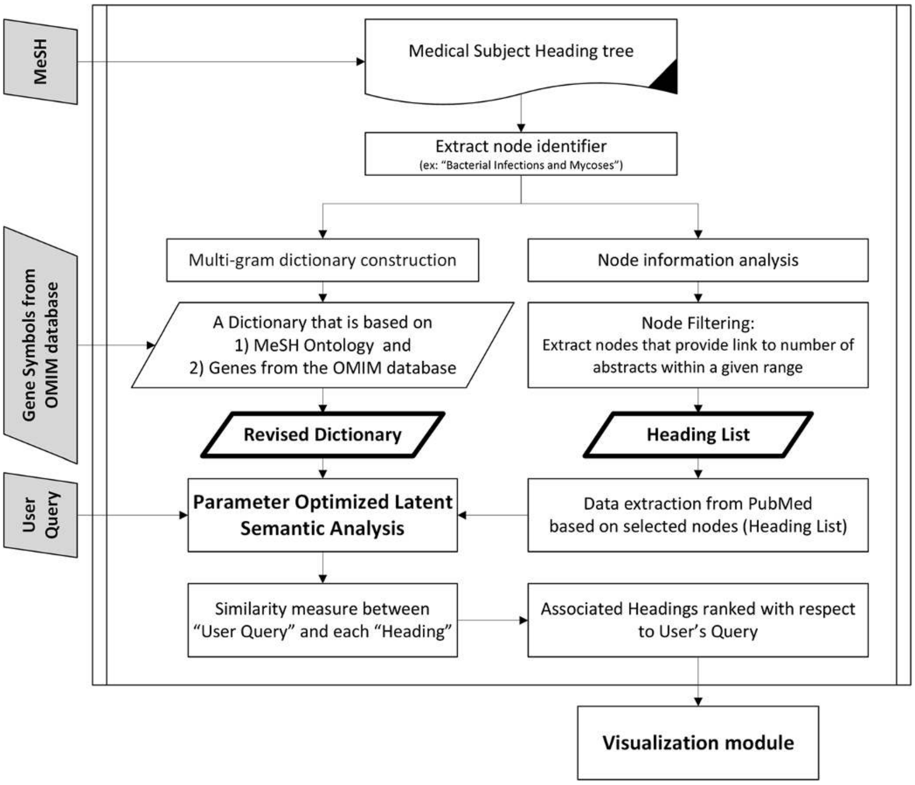

MeSH
Medical Subject Heading tree
Extract node identifier
(ex: "Bacterial Infections and Mycoses")
Gene Symbols from OMIM database
Multi-gram dictionary construction
Node information analysis
A Dictionary that is based on
1) MeSH Ontology and
2) Genes from the OMIM database
Node Filtering:
Extract nodes that provide link to number of abstracts within a given range
Revised Dictionary
Heading List
User Query
Parameter Optimized Latent Semantic Analysis
Data extraction from PubMed based on selected nodes (Heading List)
Similarity measure between "User Query" and each "Heading"
Associated Headings ranked with respect to User's Query
Visualization module

Figure 4 **ARIANA's core architecture.**

weaker yet positive associations tend to provide key indication for further in-depth investigation.

Case Study on (lethal) drug interactions in designing experiments: In 2001, an asthma research team at the John Hopkins University used the drug hexamethonium on a young healthy volunteer that ended in a tragic death due to pulmonary inflammation and fibrosis. Office for Human Research Protections of the US Department of Health and Human Services faulted the investigators for ignoring published information regarding the lung toxicity of the drug. In an internal investigation [13], the committee noted "The principal investigator subsequently stated to the investigation committee that he had performed a standard PubMed search". The committee panel referred to a number of studies, in addition to one case-report published in 1955 [14], that have reported an association between hexamethonium and pulmonary fibrosis. In that case report [14], a 28 year old woman died after receiving hexamethonium over a period of six months. Even after these two tragedies, the association between, hexamethonium and pulmonary fibrosis, or fibroma are still not evident with a keyword search from PubMed. The second tragedy was never published as a case report; nonetheless, the autopsy report as well as news broadcasts are available on the internet. This tragedy gained media's attention because it could have been prevented. In our test, ARIANA provides evidence for such associations. This knowledge was extracted even though the constructed core database contains publications from 1960 to 2012. Out of 2,545 concepts selected from the MeSH, "*Scleroderma, Systemic*", "*Neoplasms, Fibrous Tissue*", "*Pneumonia*", "*Fibroma*", and "*Pulmonary Fibrosis*" were ranked as the 13th, 16th, 38th, 174th and 257th ranked-concept, respectively. If the researchers had access to such knowledge discovery tool, capable of identifying novel associations, this investigator would likely have performed additional in-depth research before using this drug on a healthy subject. A network view of the query hexamethonium indicates that the top seven associations are relevant; however, due to the nature of the investigation, we expect the weaker associations to provide key information worth further in-depth verification by experts.

Identification of network of semantically related entities with a single or double query can uncover hidden knowledge and facilitate data reuse among other things. Alzheimer's disease (AD) is a debilitating disease of the nervous system, mostly affecting the older population. ARIANA captured some of the obvious associations such as Tauopathies; Proteostasis Deficiencies; Amyloidosis; Cerebral Arterial Diseases; Multiple System Atrophy; Agnosia. It also identified some of the less obvious associations such as Tissue Inhibitor of Metalloproteinases

[15,16]. Using Tuberculosis (TB) as a second query, a common entity was recognized to be linked to both AD and TB. "Proteostasis Deficiencies > Amyloidosis" is highly related (cosine score of 0.5651) to TB and moderately related (cosine score of 0.0734) to AD. Further investigation by expert revealed that AD and TB could be indirectly related through MMP (Matrix metalloproteinases) gene family members. MMPs are zinc-binding endopeptidases that degrade various components of the extracellular matrix [17,18]. MMPs are believed to be implicated in TB by the concept of a matrix degrading phenotype [19]. Various studies in human cells, animal models as well as gene profiling studies support the association of MMPs and TB and involvement of TB-driven lung matrix deconstruction [20-24]. MMPs are also implicated in AD [25] but in a more positive way. In fact MMP proteins can breakdown the amyloid proteins [26] that are present in the brain of the AD patients. There is literature evidence for the link between MMP genes and AD, and similarly between MMP genes and TB; however, the connection between AD and TB through the MMP genes is extracted by a global analysis of the literature.

Finally, ARIANA can be used by expert to perform global literature search using 17,074 different queries, and these include diseases, risk factors, biomedical entities and biological processes. Two additional search results from the system are summarized: 1) *Query term*: **CD4**. The five top associated headings are i) cyclin D, ii) retinal pigments > opsins, iii) human immunodeficiency virus, iv) beta-endorphin, and v) alloys > steel. 2) *Query term*: **Helicobacter pylori**. Among the top associated headings are i) apolipoproteins B, ii) adrenergic alpha-agonists, iii) isonicotinic acids, iv) oral fistula, v) identification (psychology) > gender identity and vi) diabetes mellitus, type 1. All these associated MeSH terms with the two queries have supporting evidence in the literature, even if at first some might seem unrelated. Exploring such associations, and even those that are at slightly weaker levels could provide valuable opportunities in knowledge discovery and hypothesis generation.

ARIANA is a LSA-based technique that integrates ontology mapping and advanced visualization technique to provide a global view of the knowledge that is buried in the ocean of literature. ARIANA has many advantages, such as scalability, context specificity, robustness and language-independence; however, the system has also some limitation. For instance, it is well agreed that an LSA-based technique is computationally intensive because of its utilization of Singular Value Decomposition step [7]. However, with higher computing power and the possibility to perform parallel computing this limitation can soon be overcome. A second limitation of this method is in its use of bag-of-word model, where ordering of words is lost; ARIANA uses multi-gram dictionary which alleviates this problem to some extent while still proving scalability. Finally a major different between LSA based techniques and part-of-speech tagging is LSA's inability to provide direct link to the specific publication that was the source of the identified association. We are currently working to address this specific limitation which can also be very valuable to the broader field of computational science.

Conclusion

An array of text-analytics tools [2,4,6,7] are being developed to answer and solve specific problems when dealing with biomedical literature that is increasing at an unprecedented rate. There are three main features that distinguish this work from closely related work such as Bio-LDA [4]: 1) modularity in terms of concept selection (from MeSH), 2) multi-gram dictionary construction (providing context specificity and enhanced semantics) and 3) scalability (where 50 years of literature from PubMed is analyzed). However, the system has its own limitations as stated in the discussion; our group along with others in the computational field [27-29] are actively working towards addressing these limitations..

Finally, network of semantically related associations is critical to understand the confluence between diseases, drugs, genes and risk factors. To be effective, such a tool must be efficient, robust, scalable, and useable in finding meaningful information beyond literature mining. It is the features like disambiguation of domain specific entities, flexibility in terms of visualization, broadness in coverage, robustness in modeling and scalability in providing array of Web services that made ARIANA an important tool to bridge the gap between data and knowledge.

Availability

Software is available with properly executed end users licensing agreement (EULA) at http://www.ARIANAmed.org[a].

Endnotes

[a] Requests for an account should be made to VA (abedi@vbi.vt.edu) or to MY (myeasin@memphis.edu).

[b] The list of 2,545 hierarchically-structured Headings used in the model is available upon request.

[c] The multi-gram dictionary used in the study is available upon request.

[d] http://d3js.org/.

Abbreviations

ARIANA: Adaptive robust and integrative analysis for finding novel associations; AD: Alzheimer's disease; DDD: Dynamic-data-driven; LSA: Latent semantic analysis; MeSH: Medical subject headings; MMP: Matrix metalloproteinases; OM: Ontology mapping; OMIM: Online mendelian inheritance in man; POLSA: Parameter optimized latent semantic analysis;

RM: Relevance model; ROM: Reverse ontology mapping; TB: Tuberculosis; TF-IDF: Term-frequency inverse-document-frequency.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

VA designed and carried out the experiments, and analyzed the results and drafted the manuscript. MY participated in the design of the experiments and edited the manuscript. RZ participated in the design and the analysis of the experiments and reviewed the manuscript. All authors have read, and approved the final version of the manuscript. Correspondence and requests for materials should be addressed to VA (abedi@vbi.vt.edu) or to MY (myeasin@memphis.edu).

Acknowledgments

This work was partially supported by the NSF grant NSF-IIS-0746790, Herff College of Engineering and Bioinformatics Program at the University of Memphis. The authors also acknowledge the Virginia Tech's Open Access Subvention Fund and the University of Tennessee Health Science Center (UTHSC), to cover the publication cost.

Author details

[1]The Center for Modeling Immunity to Entering Pathogens, Nutritional Immunology and Molecular Medicine Laboratory, Virginia Bioinformatics Institute, Virginia Polytechnic Institute and State University, Blacksburg, VA 24060, USA. [2]Department of Electrical and Computer Engineering, Memphis University, Memphis, TN 38152, USA. [3]College of Arts and Sciences, Bioinformatics Program, Memphis University, Memphis, TN 38152, USA. [4]Department of Neurology, University of Tennessee Health Science Center, Memphis, TN 38163, USA.

Received: 22 August 2014 Accepted: 11 November 2014
Published online: 27 November 2014

References

1. **PubMed.** [http://www.ncbi.nlm.nih.gov/pubmed]
2. Rzhetsky A, Seringhaus M, Gerstein M: **Seeking a new biology through text mining.** *Cell* 2008, **134:**9–13.
3. Wei C-H, Harris BR, Li D, Berardini TZ, Huala E, Kao H-Y, Lu Z: **Accelerating literature curation with text-mining tools: a case study of using PubTator to curate genes in PubMed abstracts.** *Database (Oxford)* 2012, **2012:**bas041.
4. Wang H, Ding Y, Tang J, Dong X, He B, Qiu J, Wild DJ: **Finding complex biological relationships in recent PubMed articles using Bio-LDA.** *PLoS One* 2011, **6:**e17243.
5. Abedi V, Zand R, Yeasin M, Faisal FE: **An automated framework for hypotheses generation using literature.** *BioData Min* 2012, **5:**13.
6. Lu Z: **PubMed and beyond: a survey of web tools for searching biomedical literature.** *Database (Oxford)* 2011, **2011:**baq036.
7. Chen H, Martin B, Daimon CM, Maudsley S: **Effective use of latent semantic indexing and computational linguistics in biological and biomedical applications.** *Front Physiol* 2013, **4:**8.
8. Landauer TK, Laham D, Derr M: **From paragraph to graph: latent semantic analysis for information visualization.** *PNAS* 2004, **101:**5214–5219.
9. Abedi V, Yeasin M, Zand R: **ARIANA: adaptive robust and integrative analysis for finding novel associations.** In *2014 Int Conf Adv Big Data Anal.* Las Vegas, NV: CSREA Press; 2014.
10. **Medical Subject Headings.** http://www.ncbi.nlm.nih.gov/mesh.
11. **Online Mendelian Inheritance in Man.** http://omim.org/.
12. Yeasin M, Malempati H, Homayouni R, Sorower M: **A systematic study on latent semantic analysis model parameters for mining biomedical literature.** *BMC Bioinformatics* 2009, **10**(Suppl 7)**:**A6.
13. Internal Investigative Committee Membership: **Report of Internal Investigation into the Death of a Volunteer Research Subject.** 2001, http://www.hopkinsmedicine.org/press/2001/july/report_of_internal_investigation.htm.
14. Robillard R, Riopelle JL, Adamkiewicz L, Tremblay G, Genest J: **Pulmonary complications during treatment with hexamethonium.** *Can Med Assoc J* 1955, **72:**448–451.
15. Wollmer MA, Papassotiropoulos A, Streffer JR, Grimaldi LME, Kapaki E, Salani G, Paraskevas GP, Maddalena A, de Quervain D, Bieber C, Umbricht D, Lemke U, Bosshardt S, Degonda N, Henke K, Hegi T, Jung HH, Pasch T, Hock C, Nitsch RM: **Genetic polymorphisms and cerebrospinal fluid levels of tissue inhibitor of metalloproteinases 1 in sporadic alzheimer's disease.** *Psychiatr Genet* 2002, **12:**155–160.
16. Ridnour LA, Dhanapal S, Hoos M, Wilson J, Lee J, Cheng RYS, Brueggemann EE, Hines HB, Wilcock DM, Vitek MP, Wink DA, Colton CA: **Nitric oxide-mediated regulation of β-amyloid clearance via alterations of MMP-9/TIMP-1.** *J Neurochem* 2012, **123:**736–749.
17. Brinckerhoff CE, Matrisian LM: **Matrix metalloproteinases: a tail of a frog that became a prince.** *Nat Rev Mol Cell Biol* 2002, **3:**207–214.
18. Davidson JM: **Biochemistry and turnover of lung interstitium.** *Eur Respir J Off J Eur Soc Clin Respir Physiol* 1990, **3:**1048–1063.
19. Elkington PT, Ugarte-Gil CA, Friedland JS: **Matrix metalloproteinases in tuberculosis.** *Eur Respir J Off J Eur Soc Clin Respir Physiol* 2011, **38:**456–464.
20. Thuong NTT, Dunstan SJ, Chau TTH, Thorsson V, Simmons CP, Quyen NTH, Thwaites GE, Lan NTN, Hibberd M, Teo YY, Seielstad M, Aderem A, Farrar JJ, Hawn TR: **Identification of tuberculosis susceptibility genes with human macrophage gene expression profiles.** *PLoS Pathog* 2008, **4**(12)**:**e1000229.
21. Mehra S, Pahar B, Dutta NK, Conerly CN, Philippi-Falkenstein K, Alvarez X, Kaushal D: **Transcriptional reprogramming in nonhuman primate (Rhesus Macaque) tuberculosis granulomas.** *PLoS One* 2010, **5**(8)**:**e122666.
22. Russell DG, VanderVen BC, Lee W, Abramovitch RB, Kim M, Homolka S, Niemann S, Rohde KH: **Mycobacterium tuberculosis wears what it eats.** *Cell Host Microbe* 2010, **8:**68–76.
23. Berry MPR, Graham CM, McNab FW, Xu Z, Bloch SAA, Oni T, Wilkinson KA, Banchereau R, Skinner J, Wilkinson RJ, Quinn C, Blankenship D, Dhawan R, Cush JJ, Mejias A, Ramilo O, Kon OM, Pascual V, Banchereau J, Chaussabel D, O'Garra A: **An interferon-inducible neutrophil-driven blood transcriptional signature in human tuberculosis.** *Nature* 2010, **466:**973–977.
24. Van der Sar AM, Spaink HP, Zakrzewska A, Bitter W, Meijer AH: **Specificity of the zebrafish host transcriptome response to acute and chronic mycobacterial infection and the role of innate and adaptive immune components.** *Mol Immunol* 2009, **46:**2317–2332.
25. Yong VW, Krekoski CA, Forsyth PA, Bell R, Edwards DR: **Matrix metalloproteinases and diseases of the CNS.** *Trends Neurosci* 1998, **21:**75–80.
26. Yan P, Hu X, Song H, Yin K, Bateman RJ, Cirrito JR, Xiao Q, Hsu FF, Turk JW, Xu J, Hsu CY, Holtzman DM, Lee J-M: **Matrix metalloproteinase-9 degrades amyloid-beta fibrils in vitro and compact plaques in situ.** *J Biol Chem* 2006, **281:**24566–24574.
27. Rusu C, Dumitrescu B: **Stagewise K-SVD to design efficient dictionaries for sparse representations.** *IEEE Signal Process Lett* 2012, **19:**631–634.
28. Yaguang D, Guofeng Z, Chenyang C, Jian Z, Liang T: **A parallel implementation of singular value decomposition based on map-reduce and PARPACK.** In *Proc 2011 Int Conf Comput Sci Netw Technol. Volume 2.* Harbin, China: IEEE; 2011:739–741.
29. Liang Z, Li W, Li Y: **A parallel probabilistic latent semantic analysis method on MapReduce platform.** In *2013 IEEE Int Conf Inf Autom.* Yinchuan, China: IEEE; 2013:1017–1022.

doi:10.1186/s12967-014-0324-9
Cite this article as: Abedi *et al.*: Empirical study using network of semantically related associations in bridging the knowledge gap. *Journal of Translational Medicine* 2014 **12**:324.