Mining scientific literature about ageing to support better understanding and treatment of degenerative diseases

Donatella Gubiani[1], Ingrid Petrič[1], Elsa Fabbretti[1] and Tanja Urbančič[1,2]

[1] University of Nova Gorica
Vipavska 13, Rožna Dolina
5000 Nova Gorica, Slovenia

[2] Jožef Stefan Institute
Jamova 39
1000 Ljubljana, Slovenia

{donatella.gubiani, ingrid.petric, elsa.fabbretti, tanja.urbancic}@ung.si

ABSTRACT

In this paper we demonstrate how literature mining can support experts in biomedicine on their way towards new discoveries. This is very important in complex, not yet sufficiently understood domains, where connections between different sub-specialities and fields of expertise have to be connected to fully understand the phenomena involved. As a case study, we present our preliminary literature mining work in the domain of ageing. The results confirm very recent discoveries about connections between diet and degenerative diseases, and indicate some concrete directions for further research needed to reveal the connections between microbiota and Alzheimer disease.

1. INTRODUCTION

Due to the rapid growing of scientific literature and the difficulties in knowledge integration between different scientific fields, IT represents a useful support to solve complex interdisciplinary questions. It has been demonstrated that connections and valuable hypotheses can be generated by linking findings across scientific literature with the use of literature mining methods and tools [5].

The first proposal of discovery based on different literatures was given by Swanson [22]. He proposed the in silico model ABC that performs a search for new indirect relations between two disjoint sets of literature. Later, two main approaches have developed [25]. The first one, namely the closed discovery process, focuses on the test of a starting hypothesis: given two starting domains *a* and *c*, and their corresponding literatures A and C, the process extracts the common terms *b*, appearing in both literatures and representing potential bridges between the domains. Differently, the open discovery process is characterized by the absence of advance specification of target concepts: starting from a specific domain *c* and the corresponding literature C, the candidates for *a* are the result of the discovery process. Some methods combine both approaches. An example is the RaJoLink method [18] that suggests candidates for *a* based on exploration of rare terms in the literature on *c*.

Literature mining has been successfully applied in the field of biomedicine. Detailed survey of the earlier work is found in [11] and [14]. Recently, Zhang et al. [26] presented an application of their literature mining tool in retrieval of comorbidities for asthma in children and adults. Oh and Deasy [19] used literature mining to investigate chemoresistance-related genes and pathways of multiple cancer types. Their comprehensive survey and analysis provide a systems biology-based overview of the underlying mechanisms of chemoresistance. Rajpal et al. [21] applied literature mining for understanding disease associations in drug discovery. They showed how a literature mining system can be used to predict emerging trends at a relatively early stage of obesity and psoriasis by analysing the literature-identified genes for genetic associations, druggability, and biological pathways. Cameron et al. [7] also implemented a literature mining method for discovering informative and potentially unknown associations between biomedical concepts. Given a pair of concepts, their method automatically generates a ranked list of subgraphs that capture multifaceted complex associations between biomedical concepts.

In our study we applied literature mining to the domain of ageing. Ageing is an urgent health priority, with social and economical implications, that requires interdisciplinary biomedical research investments to validate multi-system intervention strategies and early diagnostic and prognostic biomarkers, as well as containment tools. While single-cell mechanisms of ageing processes are studied since several years, limited knowledge is available on the changes occurring at tissue, organ and system levels leading to progression of complex chronic age-related disorders, such as cardiovascular or neuronal diseases and cancer. On the top of genetic and individual predisposition, lifestyle environment and diet strongly contribute to occurrence of ageing and age-related diseases, although link among these factors is difficult to predict.

The main contribution of the paper is the proof of principle of in silico approaches to integrate the available literature concerning "ageing". Starting from the investigation of the existing scientific literature with ontologies and exploiting the method RaJoLink, we uncover candidate hypotheses for discoveries from terms, appearing rarely in scientific literature on this topic. As these methods indicated interesting connections between dietary issues and degenerative diseases in our preliminary studies, we focus on these particular aspects.

The paper is organized as follows. Section 2 describes the general methodology used in our work. In Section 3, we describe the investigated context exploiting ontologies and, in Section 4, we summarize different steps and results by applying the method

RaJoLink. Finally, conclusions and future works are drawn in Section 5.

2. METODOLOGY

PubMed [23] is a medical bibliographic database that contains more than 24 millions citations from years '40 to the present. Starting from the selected citations, the first step of our work consisted of a systematic analysis of the considered domain using ontologies (Subsection 2.1). Then, using an open discovery process allowed to detect new hypotheses (Subsection 2.2). Since background knowledge has an important role in guiding the discovery process, we used also expert assessments on relevant scientific issues.

2.1 Ontologies and OntoGen

An ontology is a data model that represents a domain. It is used to reason about it, about its main elements and the relationships between them. Several tools have been developed to support users to construct ontologies. One of these tools is OntoGen [13], a semi-automatic and data driven ontology editor focusing on editing of topic ontologies. Starting from a set of text-documents, it combines text-mining techniques with an efficient user interface to support the construction of ontologies including concept hierarchy visualization, concepts management with suggestions based on unsupervised and supervised machine learning methods, and concept visualization (an example in Figure 1).

2.2 Open discovery process and RaJoLink

RaJoLink [18] is an associative approach to identify and connect information across different contexts (literatures). It explores the role of rare terms in the texts that are not typical for the study domain. RaJoLink involves three main steps:

- ***Ra***re: the literature about a specific problem *A* (the domain under investigation) is examined in order to identify interesting terms that rarely appear in the selected documents (*R*);
- ***Jo***int: disjoint sets of documents about the selected rare terms are inspected to detect interesting joint terms $c_1, \ldots c_n$ that appear in their intersection;
- ***Link***ing: implementing the closed discovery, the last step links terms that bridge the gap between the starting literature *A* and the literature *C*.

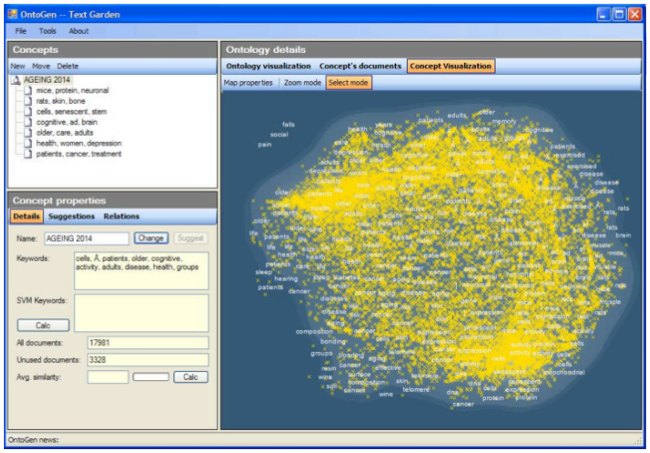

Figure 1: Concept visualization for literature about ageing in 2014.

3. ANALYSIS OF THE DOMAIN

At present, PubMed contains more than 300.000 citations about "ageing" (and its synonymous terms). When considered only the year 2014, more than 19.000 citations were found and Figure 1 shows the corresponding concept visualization. In it, we can view that recent investigations focus on different concepts, some of them connected with nutrition and food.

To validate our in silico method, we focused on important new findings, namely the link between ageing and nutrition, recently expanded by the use of high throughput -omics technologies (metagenomics, lipidomics, metabolomics, etc) and high power analysis of 'big data projects' [12]. Our work focused in a subset of 4.839 citations (with abstracts) obtained combining two properties: the conjunction of the term "ageing" with the term "food" and the most recent years (from year 2009 to 2014). Using OntoGen system, we created the two-level ontology shown in Figure 2. The first level distinguished four main data clusters. For the cluster related to the diet, at the second level we got a group of papers, related to brain.

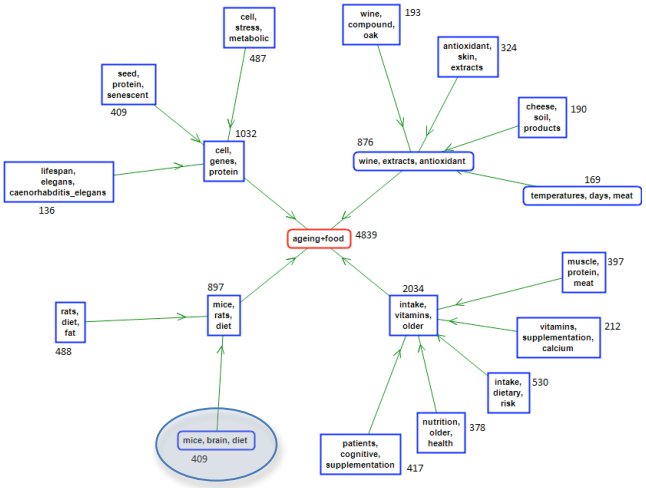

Figure 2: A two level ontology that captures the view of the domain "ageing and food".

The built ontology, obtained from the data for the last five years, revealed that "brain ageing" is a significantly reported keyword. This is consistent with the urgent need of understanding neurodegenerative diseases in elderly population. Following this route, we focused on the "Brain - Gut Axis" as described in [8] and briefly summarized at the beginning of the next section.

4. SEARCHING FOR CONNECTIONS EXPLANAINING "BRAIN-GUT AXIS"

The gut–brain axis integrates neural, hormonal and immunological signalling between the gut and the brain [8]. Bidirectional relation between these organs is mediated by neuro-active molecules, which are produced by the gut nervous system. In addition, gut microbiota biodiversity allows appropriate brain chemistry and can influence BDNF levels, learning and behavior. The diet and individual genetic and lifestyle factors influence gut microflora in their effective transformation of food in active micronutrients and essential enzymatic co-factors [4]. Interdisciplinary studies that integrated new sequencing approaches, with proteomic and metabolomic studies, have demonstrated the impact of the gut function on a wide range of health and disease processes, including chronic neuronal

disorders. A likely possibility is that personalized diet and nutrition supplements (such as probiotics) might limit chronic disease progression, cancer and ageing. These findings suggest the urgency to use IT tools to integrate data from different sectors (human health, biomedicine, microbiology, nutrition, etc) to provide guidelines for evidence based interdisciplinary research. The ability to predict a particular drug's pharmacokinetics and a given patients population's response to drugs via interfering on gut microbiome and diet, is one of the largest impacts of this field.

In the following subsection, the application of RaJoLink method and its results in this domain are described.

4.1 Rare

There is a vast interest about microbes that colonize the human gut (collectively referred as microbiota) and our health is the focus of a growing number of research initiatives. Starting from the most recent literature concerning "microbiota" (last 1000 papers with abstract available in PubMed in early June 2015), we applied the first step of RaJoLink method.

From the wide amount of rare terms detected by RaJoLink tool, we identified 3 rare terms "BDNF", "homocysteine" and "ubiquitin". While BDNF is a marker of learning abilities and brain well function [20], homocysteine is associated to ischemic brain damage [10] and ubiquitin refers to occurrence of protein post-translational modification mechanisms involved in protein quality control, a mechanism lost in neurodegenerative diseases, where large stack of unfolded unsoluble protein assembly is found [24]. Link of these terms with nutrition is subject of study, specifically associated to ageing diseases. In particular, the benefit of food supplements in endogenous BDNF levels is subject of clinical trials [9, 2].

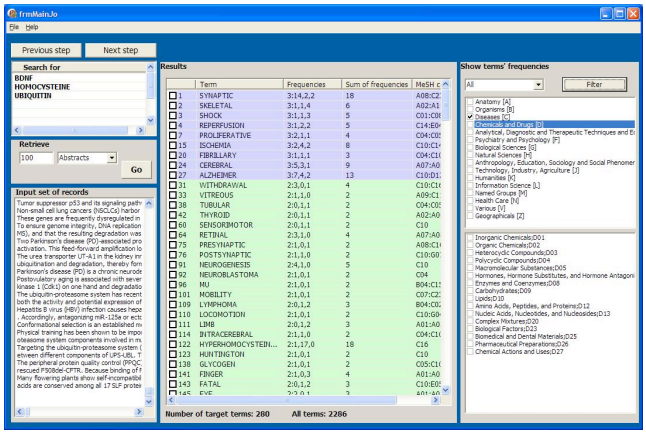

Figure 3: RaJoLink: from rare terms to candidate hypotheses.

4.2 Joint

The three rare terms "BDNF", "homocysteine" and "ubiquitin" have been the starting point for the second step in RaJoLink method. As visible in Figure 3, several joint terms have been detected. Checking their frequency, the first three terms have been "synaptic" (18 occurrences), "Alzheimer" (13 occurrences), and "cerebral" (9 occurrences).

In PubMed, we verified the connections between different literatures. In particular, Figure 4 focuses on "Alzheimer" literature: the (gut) "microbiota" and the "Alzheimer" literatures had very weak direct connection (Figure 5) but, considering joint terms and related literatures, indirect connections were found.

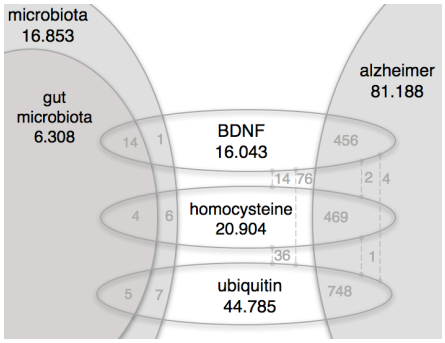

Figure 4: Analysis of connections between the scientific literature on "microbiota" and the scientific literature on Alzheimer through rare terms (PubMed: 31 August 2015).

4.3 Linking

Recent metagenomic and metabolomic screenings performed in the population, has proven the association of personal diet to individual gut microbiome profile in healthy and pathological ageing, rapidly expanding the impact of nutrition science [17]. In the next step, we performed a more detailed analysis between the scientific literatures on join terms. In particular, we focused our attention on the link between microbiota and Alzheimer literatures.

On today's date (August 2015), if we perform a search about combined literature ("microbiota" and "alzheimer"), we obtain only 9 results, 6 focusing on "gut" (Figure 5). They are published from 2013 to 2015: some of them later than papers used as input for our analysis. Significant part of these references suggests connection between the two topics being through diabetes and obesity (for gut microbiota [1, 16, 3, 6]). As described in [3], the composition of gut microbiota contributes to the development of diabetes Types 1, 2 and 3 by complex interactions of genetic and several environmental factors. Moreover, insulin resistance in the brain has been associated with Alzheimer's disease.

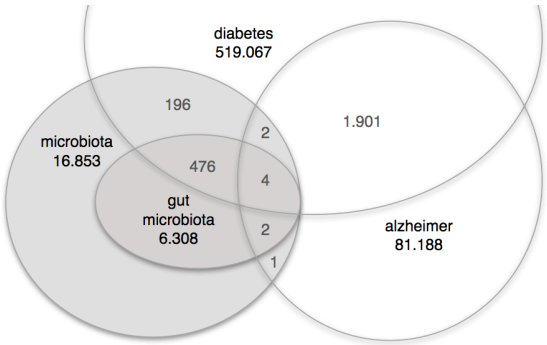

Figure 5: Connections between the scientific literature on microbiota and the scientific literature on Alzheimer (PubMed: 31 August 2015).

5. CONCLUSIONS AND FUTURE WORK

Presented literature mining methodologies are general and can be applied in different fields to guide discovery processes, providing that there is a good coverage of documents available on-line. The process is more efficient if there is a field expert available to cooperate since his or her guidance in selecting rare or joint terms might significantly fasten the convergence towards novel and interesting results. We applied literature mining to study ageing

context and we explored the role of rare terms that are not typical in the literature regarding “ageing” and “food”, to detect interesting, not yet fully understood connections, showing promising directions for further research concerning as dietary issues and degenerative diseases.

The increasing rate of data generation across all scientific fields provides new opportunities for data-driven research, with the potential to inspire new scientific trends. With the use of high throughput technologies, not only in the genetic field, but also in the metabolomic, nutrition and microbiology fields, the exploitation of 'big data' enable us to face the challenge of in silico integratomic analysis to find causative relations that might stimulate re-evaluation of existing knowledge as well as identify new data-driven medicinal chemistry and drug discovery processes [15]. In the filed of nutrition and diet, in particular, data integration analysis is often complicated by many confounding (lifestyle) factors. Validation of robust in silico tools is therefore highly desired for exploiting previous research in new interdisciplinary applications.

6. ACKNOWLEDGMENTS

This work was performed within the Creative Core project (AHA-MOMENT), partially supported by the Ministry of Education, Science and Sport, Republic of Slovenia, and European Regional Development Fund. We also acknowledge the European Commission's support through the Human Brain Project (Gr. no. 604102), as well as support of the Slovenian Research Agency through the program Knowledge Technologies and project Development and Applications of New Semantic Data Mining Methods in Life Sciences.

7. REFERENCES

[1] M. Alam, Q. Alam, M. Kamal, A. Abuzenadah, and A. Haque. A possible link of gut microbiota alteration in type 2 diabetes and Alzheimer’s disease pathogenicity: an update. *CNS Neurol Disord Drug Targets*, 13(3):383–90, 2014.

[2] J. E. Beilharz, J. Maniam, and M. J. Morris. Diet-induced cognitive deficits: The role of fat and sugar, potential mechanisms and nutritional interventions. *Nutrients*. 7(8):6719-38, 2015.

[3] P. Bekkering, I. Jafri, F. van Overveld, and G. Rijkers. The intricate association between gut microbiota and development of type 1, type 2 and type 3 diabetes. *Expert Rev Clin Immunol.*, 9(11):1031–41, 2013.

[4] J. Bienenstock, W. Kunze, and P. Forsythe. Microbiota and the gut-brain axis. *Nutr Rev.* 73 Suppl 1:28-31, 2015.

[5] P. Bruza and M. Weeber. *Literature-based Discovery*. Springer Publishing Company, Incorporated, 1 edition, 2008.

[6] R. Buchet, J. Millán, and D. Magne. Multisystemic functions of alkaline phosphatases. *Methods Mol Biol.*, 1053:27–51, 2013.

[7] D. Cameron, R. Kavuluru, T. C. Rindflesch, A. P. Sheth, K. Thirunarayan, and O. Bodenreider. Context-driven automatic subgraph creation for literature-based discovery. *Journal of Biomedical Informatics*, 54:141-57, 2015.

[8] S. M. Collins, M. Surette, and P. Bercik. The interplay between the intestinal microbiota and the brain. *Nat Rev Microbiol*., 10(11):735-42, 2012.

[9] A. D. Dangour, P. J. Whitehouse, K. Rafferty, S. A. Mitchell, L. Smith, S. Hawkesworth, and B. Vellas. B-vitamins and fatty acids in the prevention and treatment of Alzheimer's disease and dementia: a systematic review. *Journal Alzheimers Dis.* 22(1):205-24, 2010.

[10] G. Douaud, H. Refsum, C. A. de Jager, R. Jacoby, T. E. Nichols, S. M. Smith, and A.D. Smith. Preventing Alzheimer's disease-related gray matter atrophy by B-vitamin treatment. Proc Natl Acad Sci U S A, 110(23):9523-8, 2013.

[11] R. A.-A. Erhardt, R. Schneider, and C. Blaschke. Status of text-mining techniques applied to biomedical text. *Drug Discov Today*, 11 (7/8), 315-325, 2006.

[12] A. R. Ferguson, J. L. Nielson, M. H. Cragin, A. E. Bandrowski, and M. E. Martone. Big data from small data: data-sharing in the 'long tail' of neuroscience. *Nat Neurosci.*, 17(11):1442-7, 2014.

[13] B. Fortuna, D. Mladenić, and M. Grobelnik. Semi-automatic construction of topic ontologies. In *Proc. of Joint Int. Workshops on Semantics, Web and Mining*, 121–131, 2006.

[14] L. Jensen, J. Saric, and P. Bork. Literature mining for the biologist: from information retrieval to biological discovery. *Nat Rev Genet*, 7:119-129, 2006.

[15] S. J. Lusher, R. McGuire, R. C. van Schaik, C. D. Nicholson, J. de Vlieg. Data-driven medicinal chemistry in the era of big data. *Drug Discov Today*. 19(7):859-68, 2014.

[16] M. Naseer, F. Bibi, M. Alqahtani, A. Chaudhary, E. Azhar, M. Kamal, and M. Yasir. Role of gut microbiota in obesity, type 2 diabetes and alzheimer’s disease. *CNS Neurol Disord Drug Targets*, 13(2):305–11, 2014.

[17] J. K. Nicholson, E. Holmes, J. Kinross, R. Burcelin, G. Gibson, W. Jia, and S. Pettersson. Host-gut microbiota metabolic interactions. *Science*. 336(6086):1262-7, 2012.

[18] I. Petrič, T. Urbančič, B. Cestnik, and M. Macedoni-Lukšič. Literature mining method rajolink for uncovering relations between biomedical concepts. *Journal of Biomedical Informatics*, 42(2):219–227, 2009.

[19] J. H. Oh, and J. O. Deasy. A literature mining-based approach for identification of cellular pathways associated with chemoresistance in cancer. *Brief Bioinform.*, pii: bbv053, 2015.

[20] S. L. Patterson. Immune dysregulation and cognitive vulnerability in the aging brain: Interactions of microglia, IL-1β, BDNF and synaptic plasticity. Neuropharmacology. 96(Pt A):11-8, 2015.

[21] D. K. Rajpal, X. A. Qu, J. M. Freudenberg, and V. Kumar. Mining emerging biomedical literature for understanding disease associations in drug discovery. *Methods Mol Biol.*, 1159:171-206, 2014.

[22] D. R. Swanson. Medical literature as a potential source of new knowledge. *Bulletin of the Medical Library Association*, 78(1):29–37, 1990.

[23] U.S. National Library of Medicine. PubMed. url: http://www.ncbi.nlm.nih.gov/pubmed

[24] D. Vilchez, I. Saez, and A. Dillin. The role of protein clearance mechanisms in organismal ageing and age-related diseases. *Nat Commun.*, 5:5659, 2014.

[25] M. Weeber, R. Klein, and L. T. W. de Jong-van den Berg. Using concepts in literature-based discovery: Simulating Swanson’s Raynaud-fish oil and migraine-magnesium discoveries. *Journal of the American Society for Information Science and Technology*, 52(7):548–557, 2000.

[26] Y. Zhang, I. N. Sarkar, and E. S. Chen. PubMedMiner: Mining and Visualizing MeSH-based Associations in PubMed AMIA. Annu Symp Proc., 1990–1999, 2014.