Journal of Biomedical Informatics 71 (2017) 178–189

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

# Automated extraction of potential migraine biomarkers using a semantic graph

Wytze J. Vlietstra [a,*], Ronald Zielman [b], Robin M. van Dongen [b], Erik A. Schultes [c], Floris Wiesman [d], Rein Vos [a,e], Erik M. van Mulligen [a], Jan A. Kors [a]

[a] Department of Medical Informatics, Erasmus Medical Centre, Rotterdam, The Netherlands
[b] Department of Neurology, Leiden University Medical Centre, Leiden, The Netherlands
[c] Department of Human Genetics, Leiden University Medical Centre, Leiden, The Netherlands
[d] Department of Medical Informatics, Academic Medical Centre, Amsterdam, The Netherlands
[e] Department of Methodology & Statistics, Maastricht University, Maastricht, The Netherlands

ARTICLE INFO

*Article history:*
Received 29 December 2016
Revised 3 April 2017
Accepted 23 May 2017
Available online 1 June 2017

*Keywords:*
Knowledge graph
Graph semantics
Biomarker identification
Migraine biomarkers
Semantic subgraph

ABSTRACT

*Problem:* Biomedical literature and databases contain important clues for the identification of potential disease biomarkers. However, searching these enormous knowledge reservoirs and integrating findings across heterogeneous sources is costly and difficult. Here we demonstrate how semantically integrated knowledge, extracted from biomedical literature and structured databases, can be used to automatically identify potential migraine biomarkers.

*Method:* We used a knowledge graph containing more than 3.5 million biomedical concepts and 68.4 million relationships. Biochemical compound concepts were filtered and ranked by their potential as biomarkers based on their connections to a subgraph of migraine-related concepts. The ranked results were evaluated against the results of a systematic literature review that was performed manually by migraine researchers. Weight points were assigned to these reference compounds to indicate their relative importance.

*Results:* Ranked results automatically generated by the knowledge graph were highly consistent with results from the manual literature review. Out of 222 reference compounds, 163 (73%) ranked in the top 2000, with 547 out of the 644 (85%) weight points assigned to the reference compounds. For reference compounds that were not in the top of the list, an extensive error analysis has been performed. When evaluating the overall performance, we obtained a ROC-AUC of 0.974.

*Discussion:* Semantic knowledge graphs composed of information integrated from multiple and varying sources can assist researchers in identifying potential disease biomarkers.



## 1. Introduction

Biomarker identification is a costly and difficult task due to the rapid growth and fragmentation of biomedical knowledge throughout biomedical literature and numerous databases. Biomarkers are any substance, structure, or process that can be measured in the body or its products and influence or predict the incidence of outcome or disease [1]. They can be (epi)genetic, proteomic, metabolomic, viral, bacterial, and visual [2,3]. Biomarkers have many applications, including the identification of patient sub-populations, predicting drug efficacy/side effects, and monitoring disease progression, which make biomarker identification a popular and important research topic [3–5]. Three related factors make the identification of biomarkers a complex, time-consuming and knowledge intensive task. First, the continuous growth and fragmentation of knowledge simply overwhelms researchers. For example PubMed, a key biomedical literature resource, has grown from 17 million to over 23 million entries in only eight years (at an exponential growth rate of 4 percent per year) [6,7]. A similar development can be observed with the size and number of biomedical databases [8–11]. Second, potential biomarkers are often not explicitly described as such in scientific articles, especially in older literature. Often, the only information reported is that the levels of a certain biomolecule are increased or decreased in a certain disease state. Finally, biomarker identification is a task that must be repeated for different diseases.

* Corresponding author.
*E-mail address:* w.vlietstra@erasmusmc.nl (W.J. Vlietstra).

http://dx.doi.org/10.1016/j.jbi.2017.05.018
1532-0464/

Identifying biomarkers automatically using computational systems would offer researchers considerable benefits in time and effort. Such computational systems would also allow for easier replication and comparison of research results. For biomedical literature, the most important knowledge reservoir, potential biomarkers could be extracted with several literature mining techniques such as: (a) co-occurrences, where non-specific co-occurrences between compounds or genes or diseases are extracted from the literature [12,13]; (b) rule-based, where rules have to be defined manually and have a limited scope [14,15]; and (c) machine-learning techniques, which are dependent on the availability of an annotated dataset for training a classifier [16,17]. In the case of biomarker identification such a dependency on training data is contradictory: to automatically extract biomarkers using literature mining, saving time and effort, we would first need to identify and extract a smaller but representative set of biomarkers manually for the training set, effectively already reaching the goal of identifying and extracting biomarkers by spending large amounts of time and effort. Instead, the approach described in this paper is based on existing, structured knowledge represented in a knowledge graph, whose creation is not dependent on the prior availability of a training set (although a reference set is naturally required to evaluate the results of experiments afterwards). Another benefit of our approach is the possibility to include both knowledge mined from literature, as well as knowledge extracted from biomedical databases.

Our system represents knowledge as a graph composed of unique biomedical concepts and their relationships. The minimal unit of knowledge in this graph is a triple of two linked concepts and their relationship (subject – predicate – object). The sources (provenance) of each triple have also been included. By focusing the knowledge graph's representation of knowledge on individual concepts and their relationships we achieve an efficient machine actionable integration of all structured knowledge. This enables discovery of associations even when individual authors do not mention them explicitly. For example, if article A states that a particular compound is relevant for a disease, and article B states that this compound can be found in blood, an integrated representation in the knowledge graph enables automatic and speedy identification of the disease-relevant compounds found in blood.

This study aims to identify potential biochemical biomarker compounds for migraine using a knowledge graph which contains structured knowledge mined both from literature and from biomedical databases.

We chose to focus on migraine-related biomarkers for multiple reasons: (1) Migraine is a common, debilitating disease which affects millions of people worldwide. The migraine diagnosis is based on symptoms, as there are no generally accepted biomarkers for this disease [18]; (2) The pathogenesis of common migraine is largely unknown and, except for a few monogenetic subtypes, assumed to be multifactorial, which prevents us from deriving potential biomarkers based on clear causal factors such as genes or biochemical pathways [19]; (3) Migraine biomarkers are hypothesized to result in better pathophysiological understanding, improved differentiation between different headaches syndromes, prediction of treatment responses, or prediction of future chronification of this disabling disorder [5]; (4) Migraine is a well-researched disease in general, resulting in many publications, which both enables and necessitates the automated identification of potential biomarkers; (5) Computer-aided literature research into migraine has a rich history of knowledge discovery, first initiated by Swanson with his literature based discovery of the relationship between magnesium and migraine [20].

## 2. Background

Previous studies have attempted to identify and extract biomarkers from biomedical literature. Bravo et al. extracted known biomarkers by mining all literature co-occurrences between diseases and proteins or genes from Medline entries that had been annotated with the "Biological Markers" MeSH heading. They extracted 131,012 gene – disease associations, from which 11% were identified as biomarkers in DisGeNet [3]. Fleuren et al. extended the CoPub tool to CoPubGene, to create a network of gene-disease and gene-gene co-occurrences found in Medline abstracts [21]. They used CoPubGene to describe the pathophysiology underlying glucocorticoid-induced insulin resistance and to identify genetic biomarkers, and manually investigated genes suggested by their method. However, they did not label their results as true-positive or false-positive, and did not compare their results to a reference set. A drawback of both these methods is that they are based on co-occurrences, which have a lower specificity when compared to extracting triples with explicit predicates. A different approach was taken by the developers of LiverCancerMarkerRIF. They made an interface that highlights selected biomedical entities in PubMed abstracts and allows experts to annotate potential genetic biomarkers [22]. As this method relies on human annotation, it still requires extensive manual effort. A self-organizing literature mining approach was developed for the InfoCodex system, which was applied to identify diabetes and/or obesity biomarkers [23]. They report precision values ranging from 1% to 59%, and recall values of about 34% for their most reliable benchmarks. However, this self-organization is highly dependent on training data for training a classifier. KnowLife creates a knowledge graph by automatically extracting knowledge directly from literature and pharmaceutical resources such as Drugs.com, Medline, Wikipedia Health and others, with the goal of providing users the most recent information [24]. However, at the moment of writing no publications about the practical application of KnowLife exist. What all these approaches have in common is that they focus on knowledge mined from literature and do not incorporate knowledge from databases.

The Aetionomy project has developed NeuroRDF, which combines knowledge extracted from literature and databases to suggest biomarker genes for Alzheimer's disease [25]. As no reference set was available, they performed literature studies to discuss their top-ranked results, although they also did not label their results as true-positive or false-positive. In addition to the development of NeuroRDF, they performed an extensive review of available knowledge graphs which are solely based on databases [26]. Another system named Biograph was also based on knowledge extracted from databases only. The developers used 627 genes known to be associated with 29 diseases within OMIM as a reference set [27]. They achieved an AUC (area under the receiver operating characteristics (ROC) curve) of 0.861. Furthermore, they retrieved 22% of their reference set in the top 1% of their list of results. Overall, as much knowledge relevant to our task is still represented in the literature, we consider graphs which include knowledge extracted from literature to have a higher coverage.

Several companies, such as Ontotext and KNOESIS, offer semantically integrated graph databases as a commercial service [28,29]. Euretos offers a knowledge graph, which is highly similar to ours, with a workflow for biomarker identification [30]. A publically accessible knowledge graph is provided by Ontotext's LinkedLifeData, containing a large number of biomedical datasets, as well as relationships mined from Medline [31]. Drawbacks of commercial products are: (1) A lack of public availability. These are products which usually cannot be used without a (paid) license; (2) a black-box character. It is uncommon for such commercial products