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

to make their underlying software and query processes publically available; and (3) a lack of control over the integration of additional datasets, and the methodology with which these are integrated.

Existing, structured knowledge can be used for other tasks as well. For example, Kang et al. demonstrated that use of a knowledge graph for the extraction of adverse drug events from literature greatly reduced the size of the training corpus as compared to a machine-learning approach [32]. For the same task Xu and Wang showed that including existing knowledge improved the F-score by 73% when compared to an SVM methodology [33]. Pons et al. demonstrated in the recent BioCreative V challenge that the use of a knowledge graph in literature mining increases performance in compound–disease relationship extraction from literature [16].

3. Materials and methods

Our approach for the identification of potential biomarkers and the evaluation against a reference dataset is shown in Fig. 1. The input consisted of two data items and ten subsequent steps. Four steps were performed manually, with time requirements ranging from short (migraine researchers needed about one working day to define the migraine subgraph) to long (error analysis, multiple weeks). This section will further describe the two data items and the steps in the process.

3.1. Graph database

In our knowledge graph, which runs on a 1.8.3 Neo4j graph database, the 3,527,423 biomedical concepts are represented as vertices, with 68,413,238 relationships between them. The individual concepts represent units of thought, which are atomic and unique [34,35]. Two concepts can be connected to each other with one or more semantic relationships (also referred to as predicates), such as "causes" or "inhibits", thereby forming a triple.

Concepts in the graph database are based on the 2012 AA version of the Unified Medical Language System (UMLS) MetaThesaurus, which has been extended with concepts for proteins from UniProt and genes from a previously created gene dictionary [36,42]. New concepts were created for proteins and genes that could not be mapped to UMLS concepts, while the terms and identifiers for proteins and genes that could be mapped to UMLS concepts were added to their concepts. All concepts have been categorized as one or more UMLS semantic types, and a single semantic group, which are aggregations of semantic types [37,38]. The added gene and protein concepts were manually assigned to UMLS semantic types/groups.

The 171 unique relationships in the knowledge graph are based on the relationships defined in the UMLS Semantic Network, the relationships defined in the UMLS MetaThesaurus (MRREL table), and the predicates used within Semantic Medline. Our approach semantically integrates relationships extracted from Medline abstracts as provided in Semantic Medline [39] with relationships obtained from the UMLS [40], UniProt [41], the Comparative Toxicogenomics Database [43], and from the datasets contained in Linked Open Drug Data (LODD, consisting of DrugBank, DailyMed, and SIDER [44]) into a single graph database. As the reference set was the result of a structured literature review initiated in 2012, we limited ourselves to integrating dataset versions from that year.

Semantic Medline is a set of triples created by the U.S. based National Library of Medicine (who also host PubMed), by running their relationship extraction tool SemRep on sentences from Medline titles and abstracts. UniProt contains annotations about individual proteins, while the Comparative Toxicogenomics Database contains annotations about the influences of chemicals on diseases and genes, as well as the relationships between genes and diseases. The LODD datasets contain pharmacological data such as drug indications, ingredients, targets, and side effects.

The concepts and predicates in the Semantic Medline triples were already expressed in terms of our ontology and predicate thesaurus, and therefore did not require additional cross-mapping effort [44,45]. For UniProt, the Comparative Toxicogenomics Database, and the LODD databases the process for extracting relationships between concepts included the manual mapping of the implicit relations of the database schema to an explicit predicate from our predicate-thesaurus. The mapping of the subject and object from database records to the UMLS, UniProt or gene identifiers in our ontology was performed by applying our Peregrine literature mining pipeline, which matched the subject or object to a term or identifier associated with a concept [46]. An example of the mapping of information from a UniProt record to concepts in our ontology and relationships within our thesaurus is shown in Appendix A.

The resulting relationships between the concepts, extracted both from literature, database entries, and the UMLS are further enriched by including references to the sources, also referred to as the provenance of the relationships between two concepts. These relationship sources can be references to articles that describe the relationship, references to database entries, or references to the Unified Medical Language System (UMLS) Metathesaurus.

3.2. Reference set

For the reference set, a systematic literature review was performed by two migraine researchers (R.M.D. and R.Z.) in accordance with the Preferred Reporting Items for Systematic reviews and meta-Analyses (PRISMA) statement, a reporting guideline for systematic literature reviews [18]. Two searches were performed on Medline, EMBASE, and Web of Science up to August 16, 2014 for published studies on biochemical findings in (1) blood, and (2) cerebrospinal fluid (CSF) of migraine patients (see Appendix C for queries) [47]. The two migraine researchers independently assessed titles and abstracts to determine eligibility. Disagreement was resolved by discussion. Subsequently, the same researchers independently assessed full-text articles of potentially relevant studies to verify if eligibility criteria were met and to evaluate whether the results were adequately reported. Case-control studies in which one or more endogenous compounds (metabolites, peptides, proteins) were quantified in blood or CSF of migraine patients and non-migraine controls were included. An example of the extracted data required to include a compound is provided in Appendix B. An overview of the entire process has been provided in Appendix E, and results for the CSF set have been published by van Dongen et al. [18].

Not all potential biomarkers compounds identified by the systematic literature review were considered equally important. For example, while some compounds were measured in multiple studies and showed significant difference between migraine patients and controls, other compounds were only measured once in relation to migraine without a significant difference between migraine patients and controls. As an easily interpretable metric of the importance of a compound, each of them was assigned a weight equal to the number of studies where a significant difference in compound concentration in patient CSF or blood was found according to the results of the structured literature review. The weight points of the CSF and blood datasets were summed. The result was a single, integrated set of compounds, with each compound having a number of weight points as a measure of evidence. With a total of 234 studies included in the literature review, the

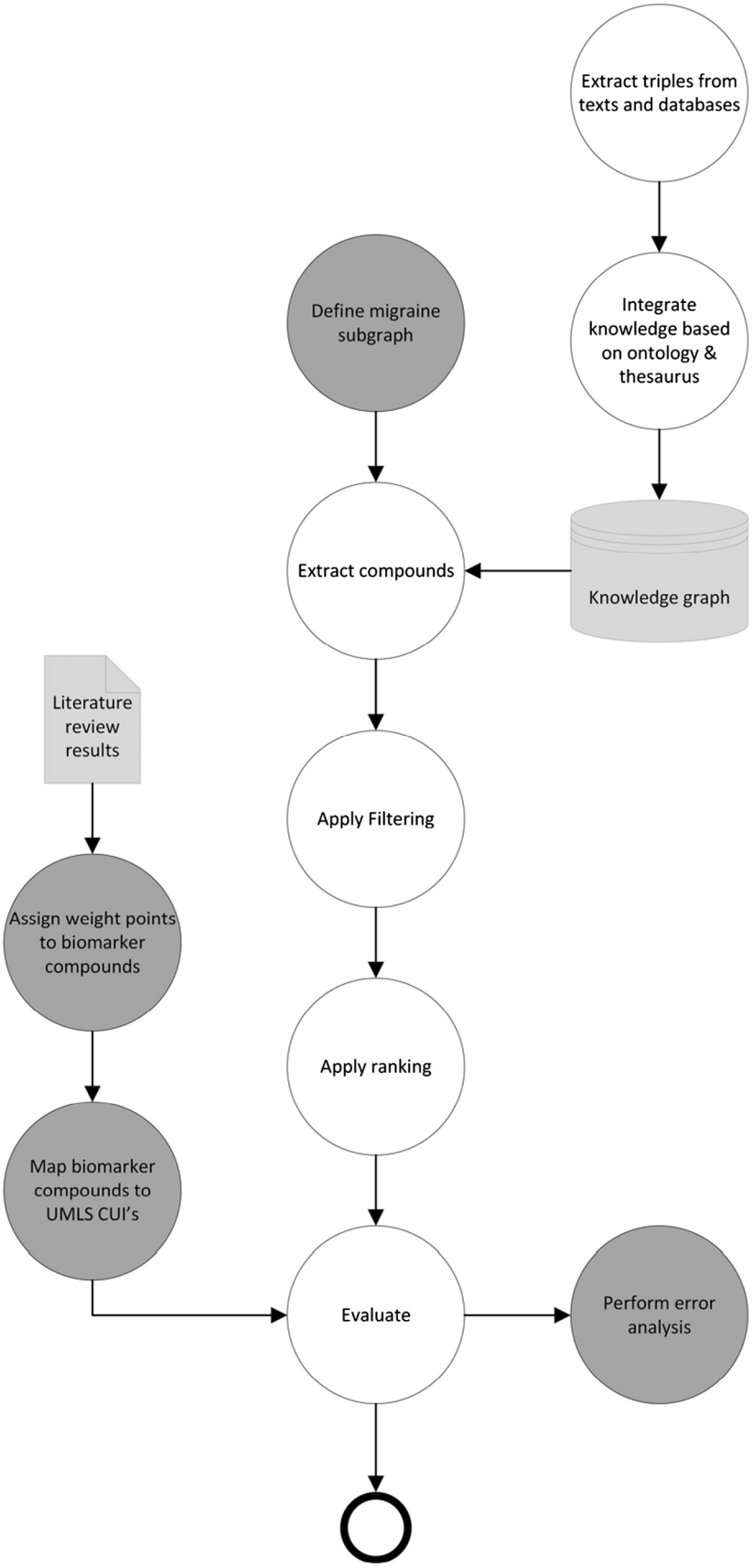


Fig. 1. Overview of the various steps in this research. The light grey items (Knowledge graph and the Literature review results) are data items which were re-used (in modified form) from other research [18,32]. Dark grey item are steps which have been performed manually. White items are steps which have been performed computationally. The empty circle at the bottom of the figure indicates the end of the process.

maximum theoretical weight for a compound was 234 points. In practice, the maximum summed weight was 25 points.

These compounds were subsequently manually mapped to UMLS identifiers. If compounds could not be exactly mapped to a term associated with a concept, a reasonable best fitting concept was selected (e.g., "Total magnesium" was mapped to "magnesium"). Some compounds that were differentiated by the migraine researchers, were considered synonymous in the UMLS (e.g., "L-arginine" & "Arginine", "Free Tryptophan" & "Tryptophan"), or not specified to the same degree (conjugated and unconjugated forms of a compound). These were considered to be duplicate entries, and one of them was removed. If duplicates differed in their weight points, the compound with the lowest number of weight points was removed. As a result of this mapping process one compound was removed from the CSF set, while ten compounds were removed from the blood set. The ultimate set of reference compounds is available in the Supplemental Materials.

Some reference compounds had one or more direct relationships with migraine, making them trivial to identify. To ensure that our method was independent of this information, it did not make use of such direct relationships between migraine and potential biomarker compounds.

3.3. The migraine subgraph

The identification of potential biomarkers was based on a migraine subgraph, where we defined a subgraph as a subset of concepts and relationships within our knowledge graph. Cameron et al. have also applied a subgraph-based methodology to recreate Swanson's original magnesium-migraine and Raynaud-fish oil discoveries based on scientific literature, which illustrates a broader trend for subgraph-based reasoning in biomedical knowledge graphs for various tasks [49]. They used subgraphs in an attempt to cluster paths consisting of specific predicates and intermediate concepts between the diseases and compounds. Our subgraph-based approach does not limit itself with such specific restrictions on predicates and intermediate concepts, and thereby enables a more holistic view of a disease where complex, cumulative interactions between disease, physiology, genetics, and chemistry can exist.

Our method searched outwards, starting from migraine concepts, then going to subgraph concepts, to end at potential biomarker compounds. A schematic representation of our model is shown in Fig. 2. Migraine concepts were all concepts that contained the word "migraine" in their label, that had a direct relationship with the primary migraine concept (UMLS identifier C0149931), and that belonged to the semantic type Disease or Syndrome. In total 58 migraine concepts adhered to these criteria. Our migraine subgraph was defined as the set of concepts that had a relationship with one of the 58 migraine concepts, and that belonged to one or more selected semantic types. Such a selection was necessary because concepts in the UMLS (and therefore the knowledge graph) range from highly specific enzymes to social behavior or laboratory procedures. The selection of the semantic types for the migraine subgraph was discussed and coordinated with the migraine researchers. An overview of the selected semantic types can be seen in Table 1. The migraine subgraph consisted of migraine-associated diseases (e.g., epilepsy), for which some of the compounds might have better described roles, as well as pathophysiological processes (e.g., cortical spreading depression) and anatomical structures [50–52]. The inclusion of the "Gene or Genome" semantic type was motivated by the genetic components of migraine, which have been extensively investigated [19]. Finally, chemical concepts with one or more relationships to this subgraph were retrieved. These chemical concepts were considered the potential biomarkers.

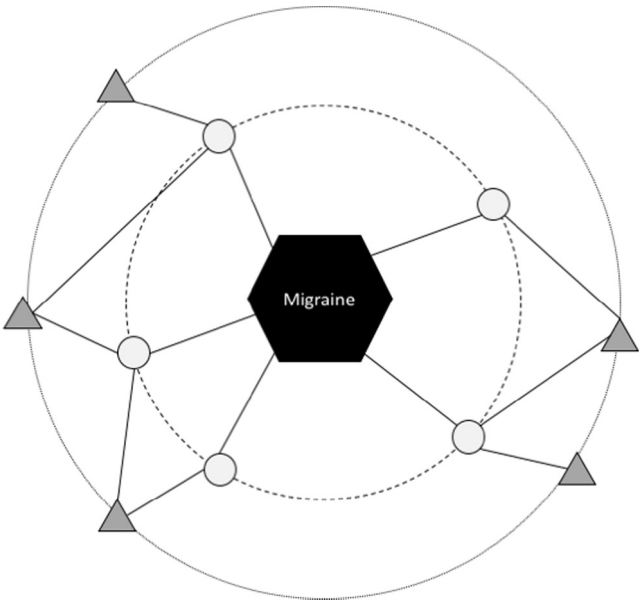


Fig. 2. Schematic representation of the migraine subgraph and potential biomarkers. The migraine subgraph consists of 58 migraine concepts (the hexagon in the centre) and migraine-related concepts (white circles). The potential biomarkers (grey triangles) are connected to the migraine subgraph.

Table 1
Overview of UMLS semantic types in the migraine subgraph.

Biologically active substance	Chemical viewed structurally
Neuroreactive substance or biogenic amine	Organic chemical
Hormone	Nucleic acid, nucleoside, or nucleotide
Enzyme	Organophosphorus compound
Vitamin	Amino acid, peptide, or protein
Immunologic factor	Carbohydrate
Receptor	Lipid
Disease or syndrome	Steroid
Mental or behavioral dysfunction	Eicosanoid
Body part, organ, or organ component	Element, ion, or isotope
Tissue	Physiologic function
Cell	Organism function
Cell component	Organ or tissue function
Gene or genome	Cell function
	Molecular function

3.4. Ranking and filtering

To improve the performance of our method we took two filtering steps, after which the potential biomarkers were ranked. This process is also represented as pseudocode in Algorithms 1 & 2.

First, we removed 24 general concepts from both the subgraph and the list of potential biomarkers, which were considered uninformative. These 24 concepts, which included concepts such as "Disease" or "Receptor" were manually identified.

Second, we removed non-endogenous compounds such as pharmaceuticals from both the subgraph and the list of potential biomarkers. Such compounds were identified by their assigned semantic types, as they have mostly been classified with the semantic type "Pharmacological Preparation", combined with a semantic type such as "Organic Chemical" or "Hormones".

Finally, we ranked the list of potential biomarker compounds based on the number of subgraph concepts they have a relationship with, as represented in Fig. 2.

Algorithm 1: Filtering and ranking process

```
#Inputs
• KG = Knowledge graph
• Subgraph = Set of migraine subgraph concepts
• Generic = Set of 24 manually selected highly generic concepts
• It is questionable whether concepts' current granularity is always necessary.
• Irrelevant = Set of excluded (combinations of) semantic types

#First filtering step
Foreach concept in Generic:
  Remove concept from Subgraph, Candidates
#Second filtering step
Foreach concept in Subgraph, Candidates:
  If semanticTypes(concept) in Irrelevant:
    Remove concept from Subgraph, Candidates

#Ranking;
Foreach concept in Candidates:
  Ranker[concept] = Count(Connections(concept,
  Subgraph, KG))
Output = Order Candidates by Ranker

Return Output
```

Algorithm 2: Connections function

```
#As our ranking method does not use predicates
  between concepts, this query function only
  returns unique combinations of subjects and
  objects.

#Inputs:
• start: A candidate biomarker concept
• end: The set of subgraph concepts
• graph: A Knowledge graph in which the subject-predicate-object triples are contained

Function Connections(start, end, graph):
  Subjects, Predicates, Objects = graph.getTriples
  (start, end)

  Return Distinct(Subjects, Objects)
```

We also experimented with three other ranking mechanisms, which are explained in Appendix D. Each of these alternative ranking mechanisms used a different level of information of the knowledge graph i.e. the edge-level, relationship-level, and provenance-level.

The resulting output of our method was a ranked list of potential biomarkers for the disease of interest. This ranking of potential biomarkers was used as an indication for their relevance.

It should be noted that our knowledge graph also contains negative predicates. When two concepts are only connected by a negative predicate, we still considered this a valid relationship and included it. Furthermore, we took no special action when two concepts were connected with contradicting predicates, as these contradictions have been found to be context dependent in 86% of the cases in Semantic Medline [53].

3.5. Evaluation method

We evaluated the results of the subgraph-based method up to k compounds with the metrics below, as it is unrealistic that users will evaluate complete result sets generated by methods such as ours [54]:

- **Recall at *k*:** The fraction of reference compounds found up to rank k.
- **Precision at *k*:** The number of reference compounds identified up to rank k, divided by k.
- **Recall of weight points at *k*:** The fraction of the total number of assigned weight points found up to k.
- **Cumulative gain:** The gain expressed in points that a user gets on average for every compound that is inspected, up to rank k. Calculated by dividing the number of the weight points found up to rank k by k.

In addition to the binary indications of the performance, such as precision and recall at k, we present a user oriented measurement of the performance with the Cumulative Gain. Not every result is equally relevant, as illustrated by the weight points assigned to the reference compounds. Järvelin and Kekäläinen have developed the cumulative gain metric, which reasons from a user perspective. Cumulative Gain quantifies the "gain" (in our case represented by the weight points) a user obtains for every inspected result. It thereby combines rank and relevance in a metric and is straightforward to interpret [55].

To evaluate the overall performance, we have plotted the commonly used ROC curve with the recall for the individual reference compounds on the Y-axis (unweighted), and a curve with the recall for the individual weight points on the Y-axis (weighted), with the unique ranking values as discrimination thresholds. For the weighted evaluation, the weights of the false-positive compounds were set to 0. For both the weighted and unweighted curves we calculated an AUC value.

4. Results

4.1. List of compounds from the systematic review

The reference set consisted of 61 CSF compounds and 200 blood compounds. The total number of weight points assigned to CSF compounds was 103 (average 1.7 per compound). The total number of weight points assigned to the blood compounds was 541 (average of 2.7 per compound). A histogram of the weight points is shown in Fig. 3. The majority of the compounds has a weight of 1, which indicates that only a single study reported their measurement. Compounds with a higher number of weight points were mostly found amongst the blood compounds.

Thirty-nine compounds were present in both subsets. Weight points assigned to these compounds had a moderate correlation

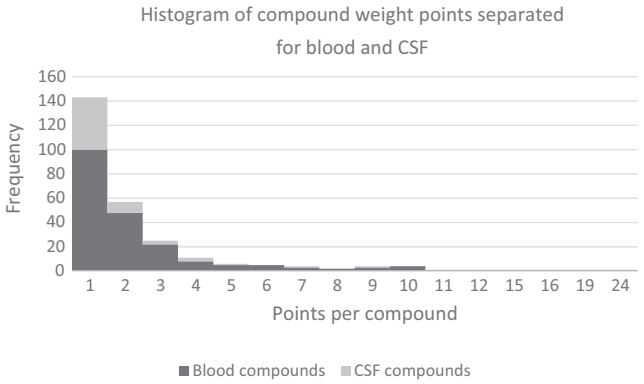


Fig. 3. Histogram of the number of weight points assigned to the compounds. The stacked bars represent CSF (grey) and blood (black).

coefficient of 0.45 (Pearson r). The two subsets were integrated into a single set of 222 compounds with a combined weight of 644 points (average of 2.9 per compound).

4.2. Migraine subgraph & potential biomarker compound list

The migraine subgraph consisted of 1060 concepts, which we categorized according to their semantic groups for brevity [37]. It consisted of 460 Disorders, 351 Chemicals & Drugs, 105 Anatomy, 99 Physiology, and 58 Genes & Molecular Sequences concepts.

The list of potential biomarkers generated by our method consisted of 51,409 extracted compounds, of which more than half had the UMLS semantic type 'Amino Acid, Peptide, or Protein', 5684 the semantic type 'Organic Chemicals', and 5543 the semantic type "Enzyme", with the remainder of the semantic types found less frequent. A complete list of all results is available in the Supplemental Materials.

4.3. Evaluation

The list of potential biomarker compounds generated by our method retrieves 201 of the 222 reference compounds (91%). We chose to cut off the table at rank 2000, and present data for various ranks of *k* up to that point, as shown in Table 2. From these 222 reference compounds, 163 reference compounds were ranked in the top 2000 of the results list (73%). These reference compounds have 547 weight points assigned to them, out of a maximum of 644 weight points (85%).

To evaluate the overall performance we calculated a ROC plot and its AUC values for both the weighted and the unweighted results, as shown in Fig. 4. We achieved an AUC value of 0.956 for the unweighted plot, and an AUC value of 0.974 for the weighted plot.

4.4. Location of reference compounds with direct relationship to migraine within the ranking

As described in Section 3.3, the method did not use direct relationships between reference compounds and migraine. From the 222 reference compounds, 87 also have a direct relationship with migraine. To establish where they were ranked, we inspected their location on the list of generated results and calculated their mean rank. The ranks and the distributions of all these compounds are shown in Fig. 5, along with their mean rank (red triangle, k = 497), and the top 1% of the 51,409 potential biomarker compounds (dark blue rhombus, k = 514). The mean rank of the reference compounds with a direct connection is within the top 1% of the list, with a considerable number of these reference compounds ranked at the very top of the list. Furthermore, the blue shape indicates that many reference compounds which only have an indirect relationship to migraine in the knowledge graph are ranked similarly high.

Table 2
Evaluation metrics calculated for the results when ranked based on the number of subgraph concepts connecting a potential biomarker compound to migraine.

k	Recall at *k*	Precision at *k*	Recall of weight points at *k*	Cumulative gain
100	0.18	0.41	0.3	1.95
200	0.31	0.35	0.5	1.6
300	0.39	0.29	0.57	1.23
400	0.46	0.26	0.65	1.04
500	0.49	0.22	0.68	0.88
600	0.52	0.19	0.7	0.76
700	0.57	0.18	0.73	0.68
800	0.6	0.17	0.75	0.6
900	0.63	0.15	0.77	0.55
1000	0.64	0.14	0.79	0.51
1100	0.66	0.13	0.79	0.46
1200	0.67	0.12	0.8	0.43
1300	0.68	0.12	0.81	0.4
1400	0.69	0.11	0.82	0.38
1500	0.69	0.1	0.82	0.35
1600	0.7	0.1	0.82	0.33
1700	0.71	0.09	0.83	0.31
1800	0.72	0.09	0.84	0.3
1900	0.73	0.09	0.85	0.29
2000	0.73	0.08	0.85	0.27

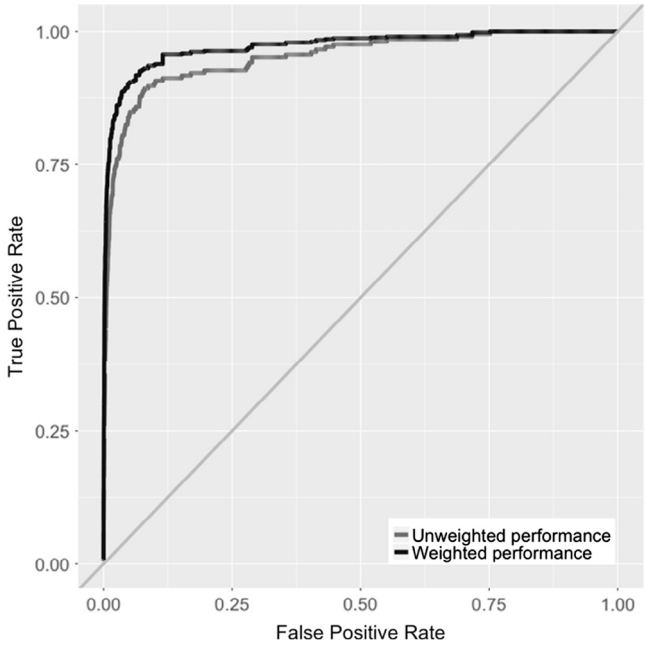


Fig. 4. ROC plot of the generated results, when ranked on the number of subgraph compounds connecting a potential biomarker compound with migraine. Both the weighted (black) and unweighted (grey) results are shown. AUC values are 0.956 for the unweighted evaluation, and 0.974 for the weighted evaluation.

4.5. Error analysis

We performed an error analysis of the false-negative compounds, i.e. the compounds that were not present within the list of results, or that were found in the middle or the tail of the result list. They were assigned to one of four error categories, as shown in Table 3. Section 4.5.3 describes an error analysis of the false-positive compounds among the top-100 compounds of the result list.

4.5.1. Not present in result list

Our method did not retrieve 21 (11%) of the reference compounds. One inclusion criterion was that compounds had to have a relationship with our migraine subgraph. For 11 reference compounds this was not the case, and they were therefore not found. Further inspection showed that these were highly specific concepts such as "6-oxo prostaglandin F1 alpha" or "para-hydroxymandelic acids (unconjugated)". For four out of these 11 compounds no relationship with another concept was available at all. The other seven compounds only had manually assigned relationships from the UMLS Metathesaurus with other concepts, but had no relationships mined from literature or databases.

As described in Section 3.3, we only included compounds with selected semantic types, and excluded non-endogenous (pharmacological) compounds. We find that 10 of the reference compounds were excluded based on these criteria.

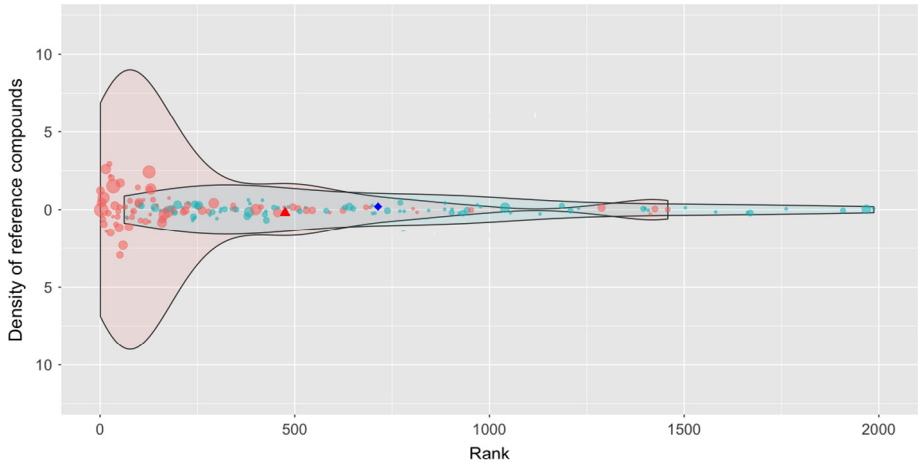


Fig. 5. Violin graph [56] of the density of reference compounds in the top 2000 compounds generated by the method. In this figure, reference compounds which also have a direct relationship with migraine (red dots) are separated from the reference compounds which only have an indirect relationship with migraine (blue dots). The dot size indicates the number of weight points assigned to the reference compound. The red and blue contours indicate the numeric density of reference compounds based on an interval of 20 ranks. The red triangle shows the mean rank of the reference compounds which also have a direct relationship with migraine (k = 497), while the dark-blue rhombus indicates 1% of the total number of generated results (k = 514). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 3
Categories of compounds we did not retrieve.

False-negatives	*n*
Not present in result list	**21**
Too far away in graph	11
Excluded during filtering	10
Low on result list	**38**
Other reference compounds	24
Long-tail compounds	14

4.5.2. Low on result list

We divided the generated list of 51,409 compounds in three groups: (1) The top 2000 compounds, (2) A group in the "middle" which we discuss below, and (3) the tail, which consisted of 37,938 compounds connected to migraine by only one or two concepts from the migraine subgraph. Fourteen reference compounds were part of the long tail. The relationships connecting them were mostly manually assigned within the UMLS Metathesaurus. Upon closer investigation, we identified two reference compounds that were ranked low due to extraction errors, likely made during the literature mining process for Semantic Medline. One of these compounds was "Orexin A" (ranked at 20,728), which was incorrectly mapped to the more general "Orexins" (which was ranked at 796) when extracted from the literature [57]. A similar error caused the relatively row ranking of "homocysteine" (ranked at 5776), which was incorrectly mapped to "homocystine" during the mining of the literature (ranked at 139) [58]. As an exhaustive investigation of such literature mining errors was beyond the scope of this article, it may well be that there were more of such occurrences.

Twenty-four reference compounds were found in the "middle" group described above, which means they were not ranked in the top 2000 results, but had multiple relationships with other concepts and had multiple sources associated with these relationships. We examined whether their lower ranking was because they were more weakly connected to the migraine subgraph, or whether this was caused by them being connected to less concepts in general. We found that for every concept in the migraine subgraph these reference compounds were connected to, they were connected to 6.46 concepts in general. We compared these compounds with the reference compounds we retrieved in the top 2000 of the results, which were connected to 7.57 concepts for every subgraph concept they were connected to. The lower ranked compounds were therefore on average slightly more specifically connected to the migraine-subgraph, by a factor 1.17. We can therefore conclude that their lower ranking was caused by them being connected to fewer concepts in general.

4.5.3. False-positives

Within the top 100 compounds, there were 59 false-positives. From these, we identified two new potential biomarker compounds which were ranked high by our method, and appeared to fulfil the requirements of the systematic literature review: "argipressin" and "estrogen" [59–61]. However, the migraine researchers considered these compounds to be the same as "vasopressin" and "estradiol", even though these are separate concepts within the UMLS [40]. Similarly, five other compounds in the top of the list were immediately recognizable as being (very) closely related to compounds from the reference set (e.g. "the human form of the TNF protein"). Another fourteen compounds should be subjected to a separate evaluation, for example by performing a knowledge discovery processes, which was beyond the scope of both the structured literature review as well as this article [20].

5. Discussion

Our research demonstrates that the complex knowledge extraction task of biomarker identification is feasible when combining structured knowledge with expert input. Additionally, the ranking of potential biomarkers can be used to rapidly provide researchers with a knowledge-based metric to prioritize biomarkers for further research. When a reference set is not used, the only user input which is required is to define the subgraph, which is a modest amount of work, although it remains to be investigated whether the methodology is generalizable for other diseases or extraction tasks.

Our subgraph-based method, which identifies potential migraine biomarkers from integrated knowledge sources, manages to identify 73% of the reference compounds, with 85% of the weight points in the top 2000 of the result list. When comparing our method's performance to BioGraph, we achieve higher ROC-AUCs, although this could be an artifact of the large number of false-positives. While the cutoff of 2000 compounds seems high, assess-

ing this number of compounds seems like a reasonable alternative for the 6435 articles the migraine researchers had to screen for their structured literature review, especially given our method's potential for further improvement such as using additional filters. As previously mentioned, the reference set only includes compounds which were measured in blood or CSF. As of yet our methodology does not include such parameters. Our initial attempts at filtering compounds, requiring them to have a relationship to CSF or blood eliminated too many reference compounds and decreased performance. By including a database such as the Human Metabolome Database (HMDB), which contains knowledge about which compounds can be found in CSF or blood, these filtering capabilities can potentially be enhanced [62].

Based on our results, the sheer number of subgraph-concepts connecting a potential biomarker compound to migraine can be used as a high performing indicator of the strength of association. While we tested another three ranking mechanisms, each using a different level of information of the knowledge graph, performance between these ranking mechanisms was roughly equal. We therefore chose to present the most intuitive and best performing ranking mechanism in the article.

Our method failed to retrieve 21 of the 222 reference compounds, while 38 compounds were ranked so low they were practically impossible to retrieve. For 6 of these 38 compounds the chosen ranking mechanism is decisive whether they are in the top 2000 results or not. For 14 out of these 38, the only associated source was the UMLS Metathesaurus, making their retrieval on the result list unrealistic. Furthermore, we showed that for at least two of these compounds their low ranks were caused by extraction errors, likely to be caused by mapping errors during the literature mining process. For example, the false-positive concept "Orexins" was ranked highly, while the true-positive "Orexin A" was ranked in the tail of the list of results. Some false-positive concepts at high ranks, such as "the human form of TNF-alpha", were recognized to be closely related to reference compounds. In two other cases the migraine researchers considered two compounds synonymous, even though this was not immediately obvious to a non-expert, e.g. "Vasopressin" and "Argipressin". Given these findings, it is unclear whether the current granularity of the results provided by the concepts is truly a requirement for users. It may be possible to collapse closely related concepts with each other to condense result lists. Such condensation of results would associate closely related concepts with each other, thereby requiring users to inspect fewer results without compromising the goal of their extraction task. The term "scientific lens" has been proposed for this approach, which allows different granularities of results for different user requirements, e.g. when it may not be necessary to specify the conjugation status of a compound [63]. As future research, we intend to investigate the impact of these scientific lenses on knowledge extraction tasks. Finally, although we use the reference set as a gold standard, it is probably not perfect. Hence, it is conceivable that some of the retrieved compounds not present in the reference set are nonetheless potential biomarkers for migraine.

When assessing the direct relationships between migraine and the reference compounds in the knowledge graph, we found them to be incomplete. Less than half of the reference compounds had a relationship with the migraine concepts, while all reference compounds have been described as being related with migraine in the literature. This discrepancy may have several causes: (1) insufficient annotations in databases, or the limitations of the literature mining process. While for the manual process knowledge extraction was based on full-text articles, including tables and figures, our process was based on knowledge extracted from individual sentences in article abstracts and database annotations, which contain much less information [48,64]; (2) not all articles used as sources in the structured literature review were found in Medline, as shown in Appendix C; (3) Besides these technical issues, we must consider the cognitive processes of the migraine researchers when they were reading the articles for the literature review. They might be able to summarize complex (transitive) relationships required for biomarker identification into more direct relationships, whereas a literature mining process separates such steps (e.g. "A has a relationship with B, and B has a relationship with C, therefore A and C may have a meaningful relationship" can be determined by a human reader, but such meaningful inferences are nontrivial to perform computationally). This creates a more complex representation of the knowledge than the cognitive model of the migraine researchers.

As future work we intend to investigate whether we can identify and leverage existing clusters of highly interconnected concepts within the knowledge graph to use as more natural subgraphs for these kinds of tasks. Furthermore, the influence of scientific development and additional knowledge has not yet been quantified. We therefore intend to repeat the experiment with more recent knowledge, up until the last iteration of the systematic literature review on 16 August 2014, and with more databases integrated into the graph to measure how much results will differ. This will allow us to quantify the value of additional knowledge for knowledge extraction tasks.

Conflict of interest

All authors declare no conflict of interest.

Appendix A. Example of mapping database record to triples

The challenge of integrating UniProt entries lies in mapping the annotation fields to the corresponding ontology concepts. We used our concept identification pipeline Peregrine to identify concepts in the free text UniProt annotation fields [46]. The mapping of the implicit relations defined in the UniProt keywords to the proper semantic predicates is a one-time manual effort and requires understanding of the biological meaning of the data. This mapping process has been performed for all integrated databases. An example of such mappings can be seen in Table 4. Once created a mapping can be applied to each update of the database. The UniProt record also contains references to the sources for these annotations (such as PubMed ID's), which have been omitted from the example table.

Appendix B. Example of source data in manual systematic review

"Results: *We assessed plasma samples from 103 women with CM [Chronic Migraine], 31 matched healthy women, 43 matched women with EM [Episodic Migraine], and 14 patients with episodic cluster headache matched for age. CGRP levels were significantly increased in CM (74.90 pg/mL) as compared with control healthy women (33.74 pg/mL), women with EM (46.37 pg/mL), and patients with episodic cluster headache (45.87 pg/mL).*" Quote taken from Cernuda-Morollón et al. [65].

Appendix C. Search strings for PubMed, EMBASE and Web of Science

C.1. PubMed

C.1.1. Migraine AND Cerebrospinal fluid

("Migraine Disorders"[Mesh] OR "migraine"[all fields] OR "Sick Headache"[all fields] OR "Migraineurs"[all fields] OR "Migraineur" [all fields] OR "migrain*"[all fields]) AND ("Cerebrospinal Fluid"

Table 4
A mapping of some of the UniProt record fields for 14-3-3 protein beta/alpha to an RDF triple. This protein is mapped from the subject resource http://www.uniprot.org/uniprot/P31946. The UniProt Keyword is manually mapped to the closest RDF thesaurus predicate. The Annotation field contents have been mapped to ontology concepts using the Peregrine concept identification pipeline.

UniProt keyword	Keyword mapped to predicate	UniProt annotation	Annotation mapped to object
gene	gene_product_encoded_by_gene	YWHAB	UMLS C1421558
GO - Molecular function	gene_product_has_biochemical_function	enzyme binding	UMLS C1149286
GO - Molecular function	gene_product_has_biochemical_function	histone deacetylase binding	UMLS C1323310
GO - Biological process	gene_product_plays_role_in_biological_process	activation of MAPKK activity	UMLS C1155556
GO - Biological process	gene_product_plays_role_in_biological_process	epidermal growth factor receptor signalling pathway	UMLS C1155379
Keywords - Biological process	gene_product_plays_role_in_biological_process	Host-virus interaction	UMLS C0599952
Subcellular location	location_of	Cytoplasm	UMLS C0010834
Keywords - Cellular component	part_of	Cytoplasm	UMLS C0010834
Keywords - Cellular component	part_of	perinuclear region of cytoplasm	UMLS C2253855
Keywords - Coding sequence diversity	gene_product_has_abnormality	Polymorphism	UMLS C0032529
Organism	conceptual_part_of	Homo sapiens (Human)	UMLS C0086418

[Mesh] OR "Cerebrospinal Fluid"[all fields] OR "Cerebrospinal Fluids"[all fields] OR "CSF"[all fields]).

C.1.2. Migraine AND Plasma/Serum

("Migraine Disorders"[Mesh] OR "migraine"[all fields] OR "Sick Headache"[all fields] OR "Migraineurs"[all fields] OR "Migraineur"[all fields] OR "migrain*"[all fields]) AND ("Plasma"[Mesh] OR "Plasma"[all fields] OR "Plasmas"[all fields] OR "Plasm"[all fields] OR "Serum"[Mesh] OR "Serum"[all fields] OR "Sera"[all fields] OR "Serums"[all fields] OR "Serologic"[all fields]).

C.2. EMBASE

C.2.1. Migraine AND Cerebrospinal fluid

(exp "Migraine"/OR "migraine".mp OR "Sick Headache".mp OR "Migraineurs".mp OR "Migraineur".mp OR "Migrain*".mp) AND (Cerebrospinal Fluid/OR Cerebrospinal Fluid.mp OR Cerebrospinal Fluids.mp OR CSF.mp).

C.2.2. Migraine AND Plasma/Serum

(exp "Migraine"/OR "migraine".mp OR "Sick Headache".mp OR "Migraineurs".mp OR "Migraineur".mp OR "Migrain*".mp) AND ((exp "Plasma"/ OR "Plasma".mp OR "Plasmas".mp OR "Plasm".mp) OR (exp "Serum"/ OR "Serum".mp OR "Sera".mp OR "Serums".mp OR "Serologic".mp)).

C.3. Web of science

C.3.1. Migraine AND Cerebrospinal fluid

TS = (Migraine OR Sick Headache OR Migraineurs OR Migraineur OR Migrain*) AND TS = (Cerebrospinal Fluid OR Cerebrospinal Fluids OR CSF).

C.3.2. Migraine AND Plasma/Serum

TS = (Migraine OR Sick Headache OR Migraineurs OR Migraineur OR Migrain*) AND TS = ("Plasma" OR "Plasmas" OR "Plasm" OR "Serum" OR "Sera" OR "Serums" OR "Serologic").

Appendix D. Alternative ranking mechanisms

In addition to ranking potential biomarkers based on the number of subgraph concepts it was connected to, we examine three levels of information contained in the knowledge graph: (1) Edge level, (2) Relationship level, (3) Source level.

An edge shared between two concepts can represent multiple kinds of relationships (i.e. predicates) between concepts. Edges thereby represent the generalization "has_semantic_relationship_with" of the specific predicates, and no two concepts are connected by more than one edge. For example, both the predicate "associated with" and "causes" between two concepts are represented by the same edge. The second level includes the explicit relationships between two concepts such as "causes" or "stimulates" and is therefore already more specific. The third level, which consists of the sources underlying the relationships, is used similarly as the second level.

Once all indirect paths between the migraine concept and the potential biomarker have been retrieved, we calculate multiple statistics to rank the potential biomarkers. These statistics are based on two-step probability calculations, each using different information from within these paths. We calculate the probability through the migraine subgraph (step 1) to a potential biomarker (step 2). The potential biomarkers are finally ranked based on the combined chance.

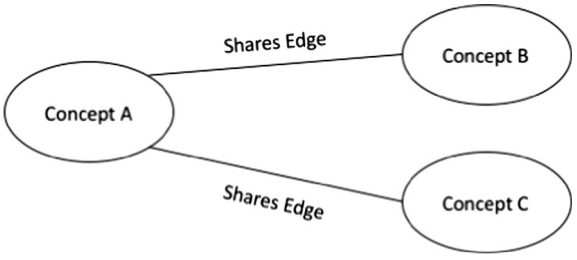


Fig. 6. Probability calculation based on edges between concepts. Here concept A shares an edge with concepts B and C. If a traversal would be started at concept A, the chance would be 0.5 for it to end up for both concept B and C.

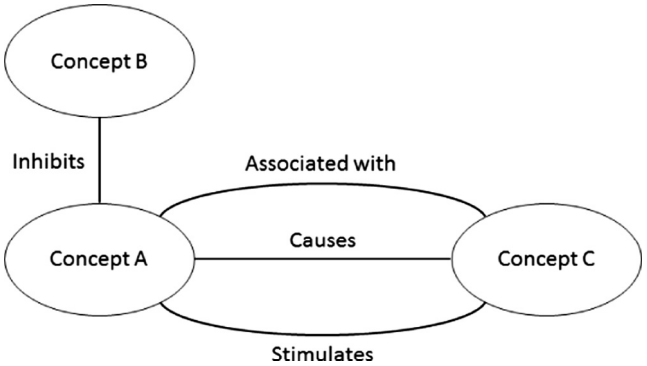


Fig. 7. Probability calculation based on relationships between concepts. Here concept A has one relationship with concept B, and three relationships with concept C. If a traversal would be started at concept A, the chance of it ending at concept B be would be 1/4 = 0.25, and 3/4 = 0.75 of it ending at concept C.

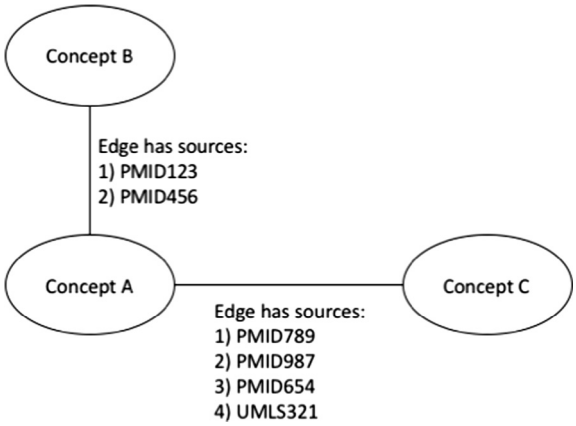


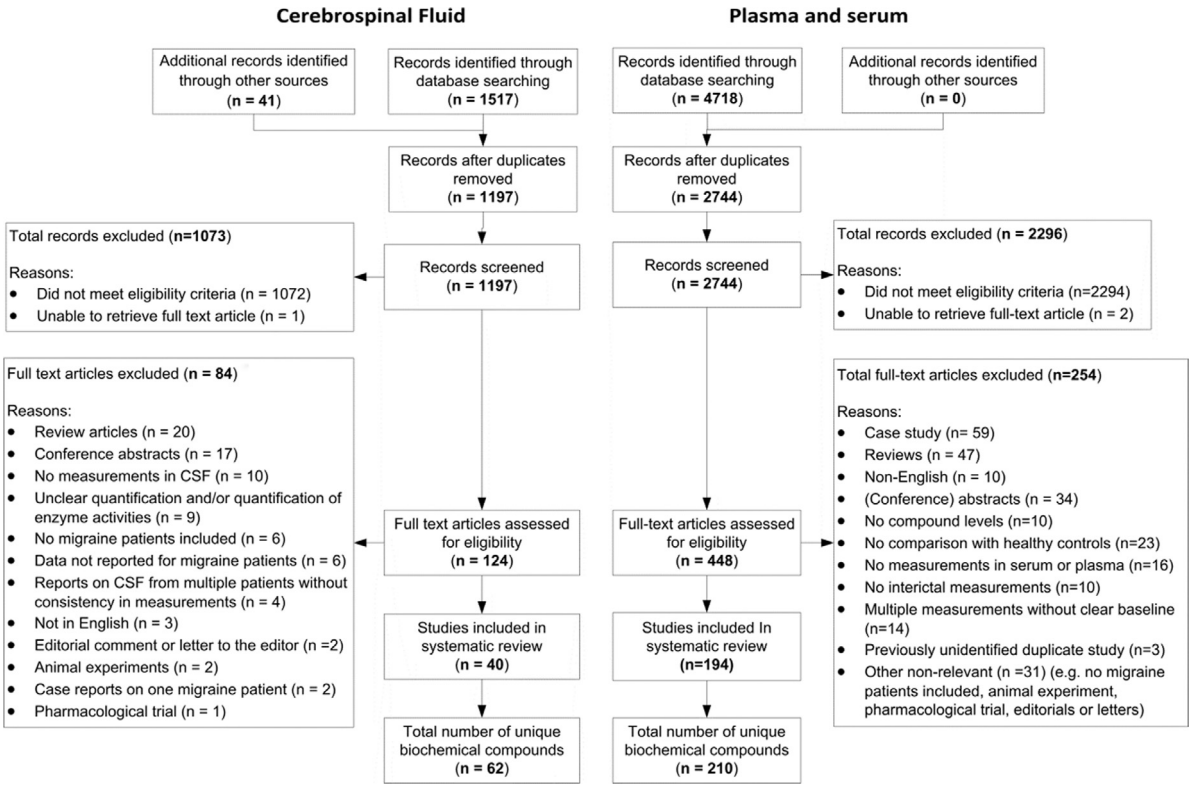
Fig. 8. Probability calculation based on sources connecting concepts. Here concept A has two sources connecting it with concept B, and 4 sources connecting it with concept C. If traversal would be started at concept A, the chance of it ending at concept B be 2/6 = 0.33, and 4/6 = 0.66 of it ending at concept C.

A formula to calculate the ultimate chance would looks as follows:

Potential biomarker relevancy
= sum(chance of step to subgraph concept)
∗ sum(chance of step to potential biomarker)

(see Figs. 6–8)).

Appendix E. Flowchart of systematic review process



Appendix F. Supplementary material

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.jbi.2017.05.018.

References

[1] WHO Biomarker, (n.d.). http://www.inchem.org/documents/ehc/ehc/ehc222.htm (accessed December 15, 2015).
[2] A. Mishra, M. Verma, Cancer biomarkers: are we ready for the prime time? Cancers (Basel) 2 (2010) 190–208, http://dx.doi.org/10.3390/cancers2010190.
[3] A. Bravo, M. Cases, N. Queralt-Rosinach, F. Sanz, L.I. Furlong, A knowledge-driven approach to extract disease-related biomarkers from the literature, Biomed Res. Int. 2014 (2014), http://dx.doi.org/10.1155/2014/253128.
[4] F. Goodsaid, Challenges of biomarkers in drug discovery and development, Expert Opin. Drug Discov. 7 (2012) 457–461, http://dx.doi.org/10.1517/17460441.2012.679615.
[5] E. Loder, P. Rizzoli, Biomarkers in migraine: their promise, problems, and practical applications, Headache 46 (2006) 1046–1058, http://dx.doi.org/10.1111/j.1526-4610.2006.00498.x.
[6] Z. Lu, PubMed and beyond: a survey of web tools for searching biomedical literature, Database 2011 (2011) baq036, http://dx.doi.org/10.1093/database/baq036.
[7] K.Z. Vardakas, G. Tsopanakis, A. Poulopoulou, M.E. Falagas, An analysis of factors contributing to PubMed's growth, J. Informetr. 9 (2015) 592–617, http://dx.doi.org/10.1016/j.joi.2015.06.001.
[8] X.M. Fernández-Suárez, M.Y. Galperin, The nucleic acids research database issue and the online molecular biology database collection, Nucleic Acids Res. 41 (2013) 1–7, http://dx.doi.org/10.1093/nar/gks1297.
[9] X.M. Fernández-Suárez, D.J. Rigden, M.Y. Galperin, The nucleic acids research database issue and an updated NAR online molecular biology database collection, Nucleic Acids Res. 42 (2014) 1–6, http://dx.doi.org/10.1093/nar/gkt1282.
[10] M.Y. Galperin, D.J. Rigden, X.M. Fernández-Suárez, The 2015 nucleic acids research database issue and molecular biology database collection, Nucleic Acids Res. 43 (2015) D1–D5, http://dx.doi.org/10.1093/nar/gku1241.

[11] D.J. Rigden, X.M. Fernandez-Suarez, M.Y. Galperin, The database issue of Nucleic Acids Research and an updated molecular biology database collection, Nucleic Acids Res. 44 (2016) D1–D6, http://dx.doi.org/10.1093/nar/gkv1356.

[12] N.R. Smalheiser, D.R. Swanson, Using ARROWSMITH: a computer-assisted approach to formulating and assessing scientific hypotheses, Comput. Methods Programs Biomed. 57 (1998) 149–153, http://dx.doi.org/10.1016/S0169-2607(98)00033-9.

[13] J. Preiss, M. Stevenson, R. Gaizauskas, Exploring relation types for literature-based discovery, J. Am. Med. Inform. Assoc. 44 (2015) ocv002, http://dx.doi.org/10.1093/jamia/ocv002.

[14] T.C. Rindflesch, M. Fiszman, The interaction of domain knowledge and linguistic structure in natural language processing: interpreting hypernymic propositions in biomedical text, J. Biomed. Inform. 36 (2003) 462–477, http://dx.doi.org/10.1016/j.jbi.2003.11.003.

[15] A. Ben Abacha, P. Zweigenbaum, Automatic extraction of semantic relations between medical entities: a rule based approach, J. Biomed. Semantics 2 (Suppl 5) (2011) S4, http://dx.doi.org/10.1186/2041-1480-2-S5-S4.

[16] E. Pons, S.A. Akhondi, Z. Afzal, E.M. Van Mulligen, J.A. Kors, RELigator: chemical-disease relation extraction using prior knowledge and textual information, BioCreative V. (n.d.) 247–253.

[17] B. Rink, S. Harabagiu, K. Roberts, Automatic extraction of relations between medical concepts in clinical texts, J. Am. Med. Infor. Assoc. 18 (2011) 594–600, http://dx.doi.org/10.1136/amiajnl-2011-000153.

[18] R.M. van Dongen, R. Zielman, M. Noga, O.M. Dekkers, T. Hankemeier, A.M. van den Maagdenberg, et al., Migraine biomarkers in cerebrospinal fluid: a systematic review and meta-analysis, Cephalalgia. (2016) 1–15, http://dx.doi.org/10.1177/0333102415625614.

[19] B. de Vries, R.R. Frants, M.D. Ferrari, A.M.J.M. van den Maagdenberg, Molecular genetics of migraine, Hum. Genet. 126 (2009) 115–132, http://dx.doi.org/10.1007/s00439-009-0684-z.

[20] D.R. Swanson, Migraine and magnesium – eleven neglected connections.pdf, Perspect Biol. Med. 31 (1988) 526–557, http://dx.doi.org/10.1353/pbm.1988.0009.

[21] W.W. Fleuren, E.J. Toonen, S. Verhoeven, R. Frijters, T. Hulsen, T. Rullmann, et al., Identification of new biomarker candidates for glucocorticoid induced insulin resistance using literature mining, BioData Min. 6 (2013) 2, http://dx.doi.org/10.1186/1756-0381-6-2.

[22] W.L. Hsu, LiverCancerMarkerRIF: a liver cancer biomarker interactive curation system combining text mining and expert annotations, Database (Oxford) 2014 (2014) 1–11, http://dx.doi.org/10.1093/database/bau085.

[23] C.A. Trugenberger, C. Wälti, D. Peregrim, M.E. Sharp, S. Bureeva, Discovery of novel biomarkers and phenotypes by semantic technologies, BMC Bioinform. 14 (2013) 51, http://dx.doi.org/10.1186/1471-2105-14-51.

[24] P. Ernst, A. Siu, G. Weikum, KnowLife: a versatile approach for constructing a large knowledge graph for biomedical sciences, BMC Bioinform. 16 (2015) 157, http://dx.doi.org/10.1186/s12859-015-0549-5.

[25] A. Iyappan, S.B. Kawalia, T. Raschka, M. Hofmann-Apitius, P. Senger, NeuroRDF: semantic integration of highly curated data to prioritize biomarker candidates in Alzheimer's disease, J. Biomed. Semantics. 7 (2016) 45, http://dx.doi.org/10.1186/s13326-016-0079-8.

[26] M. Hofmann-Apitius, G. Ball, S. Gebel, S. Bagewadi, B. De Bono, R. Schneider, et al., Bioinformatics mining and modeling methods for the identification of disease mechanisms in neurodegenerative disorders, Int. J. Mol. Sci. 16 (2015) 29179–29206, http://dx.doi.org/10.3390/ijms161226148.

[27] A.M. Liekens, J. De Knijf, W. Daelemans, B. Goethals, P. De Rijk, J. Del-Favero, BioGraph: unsupervised biomedical knowledge discovery via automated hypothesis generation, Genome Biol. 12 (2011) R57, http://dx.doi.org/10.1186/gb-2011-12-6-r57.

[28] Ontotext URL, (n.d.). http://ontotext.com/company/customers/astrazeneca-causality-data-mining-linked-data/ (accessed March 17, 2016).

[29] KNOESIS, (n.d.). http://knoesis.org/research/bionformatics (accessed March 17, 2016).

[30] Euretos Biomarkers, (n.d.). http://www.euretos.com/knowledge-platform/diagnostic-biomarker-identification (accessed September 22, 2015).

[31] LinkedLifeData, (n.d.). http://linkedlifedata.com/ (accessed March 31, 2016).

[32] N. Kang, B. Singh, C. Bui, Z. Afzal, E.M. van Mulligen, J.A. Kors, Knowledge-based extraction of adverse drug events from biomedical text, BMC Bioinform. 15 (2014) 64, http://dx.doi.org/10.1186/1471-2105-15-64.

[33] R. Xu, Q. Wang, Comparing a knowledge-driven approach to a supervised machine learning approach in large-scale extraction of drug-side effect relationships from free-text biomedical literature, BMC Bioinform. 16 (2015) S6, http://dx.doi.org/10.1186/1471-2105-16-S5-S6.

[34] P. Groth, A. Gibson, J. Velterop, The anatomy of a nano-publication, Inf. Serv. Use – Sel. Pap. From ICSTI Interact. Publ. Conf. 2010 30 (2010) 51–56.

[35] NLM UMLS, (n.d.). https://www.nlm.nih.gov/research/umls/new_users/glossary.html (accessed November 26, 2015).

[36] J. Kors, M. Schuemie, B. Schijvenaars, M. Weeber, B. Mons, Combination of genetic databases for improving identification of genes and proteins in text, BioLINK, Detroit, 2005.

[37] NLM Semantic Network, (n.d.). http://www.nlm.nih.gov/pubs/factsheets/umlssemn.html (accessed October 5, 2015).

[38] A.T. McCray, A. Burgun, O. Bodenreider, Aggregating UMLS semantic types for reducing conceptual complexity, Stud. Health Technol. Inform. 84 (2001) 216–220, http://dx.doi.org/10.3233/978-1-60750-928-8-216.

[39] H. Kilicoglu, D. Shin, M. Fiszman, G. Rosemblat, T.C. Rindflesch, SemMedDB: a PubMed-scale repository of biomedical semantic predications, Bioinformatics 28 (2012) 3158–3160, http://dx.doi.org/10.1093/bioinformatics/bts591.

[40] O. Bodenreider, The Unified Medical Language System (UMLS): integrating biomedical terminology, Nucleic Acids Res. 32 (2004) 267D–270D, http://dx.doi.org/10.1093/nar/gkh061.

[41] The UniProt Consortium, UniProt: a hub for protein information, Nucleic Acids Res. 43 (2015) D204–D212, http://dx.doi.org/10.1093/nar/gku989.

[42] D. Maglott, J. Ostell, K.D. Pruitt, T. Tatusova, Entrez gene: gene-centered information at NCBI, Nucleic Acids Res. 39 (2011), http://dx.doi.org/10.1093/nar/gkq1237.

[43] A.P. Davis, C.G. Murphy, R. Johnson, J.M. Lay, K. Lennon-Hopkins, C. Saraceni-Richards, et al., The comparative toxicogenomics database: update 2013, Nucleic Acids Res. 41 (2013), http://dx.doi.org/10.1093/nar/gks994.

[44] M. Samwald, A. Jentzsch, C. Bouton, C. Kallesøe, E. Willighagen, J. Hajagos, et al., Linked open drug data for pharmaceutical research and development, J. Cheminform. 3 (2011) 19, http://dx.doi.org/10.1186/1758-2946-3-19.

[45] Semantic Medline, (n.d.). https://skr3.nlm.nih.gov/SemMed/ (accessed May 10, 2016).

[46] M.J. Schuemie, R. Jelier, J.A. Kors, Peregrine: lightweight gene name normalization by dictionary lookup, Proc. Second BioCreative Chall. Eval. Work (2007) 131–133.

[47] A. Liberati, D.G. Altman, J. Tetzlaff, C. Mulrow, P.C. Gøtzsche, J.P.A. Ioannidis, et al., The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: explanation and elaboration, PLoS Med. 6 (2009) e1000100, http://dx.doi.org/10.1371/journal.pmed.1000100.

[48] D. Cameron, O. Bodenreider, H. Yalamanchili, T. Danh, S. Vallabhaneni, K. Thirunarayan, et al., A graph-based recovery and decomposition of Swanson's hypothesis using semantic predications, J. Biomed. Inform. 46 (2013) 238–251, http://dx.doi.org/10.1016/j.jbi.2012.09.004.

[49] D. Cameron, R. Kavuluru, T.C. Rindflesch, A.P. Sheth, K. Thirunarayan, O. Bodenreider, Context-driven automatic subgraph creation for literature-based discovery, J. Biomed. Inform. 54 (2015) 141–157, http://dx.doi.org/10.1016/j.jbi.2015.01.014.

[50] D. Pietrobon, M.A. Moskowitz, Pathophysiology of migraine, Annu. Rev. Physiol. 75 (2013) 365–391, http://dx.doi.org/10.1146/annurev-physiol-030212-183717.

[51] R. Burstein, R. Noseda, D. Borsook, Migraine: multiple processes, complex pathophysiology, J. Neurosci. 35 (2015) 6619–6629, http://dx.doi.org/10.1523/JNEUROSCI.0373-15.2015.

[52] J.P. Dreier, The role of spreading depression, spreading depolarization and spreading ischemia in neurological disease, Nat. Med. 17 (2011) 439–447, http://dx.doi.org/10.1038/nm.2333.

[53] S. Yoon, J. Jung, H. Yu, M. Kwon, S. Choo, K. Park, et al., Context-based resolution of semantic conflicts in biological pathways, BMC Med. Inform. Decis. Mak. 15 (2015) S3, http://dx.doi.org/10.1186/1472-6947-15-S1-S3.

[54] M. Yetisgen-Yildiz, W. Pratt, A new evaluation methodology for literature-based discovery systems, J. Biomed. Inform. 42 (2009) 633–643, http://dx.doi.org/10.1016/j.jbi.2008.12.001.

[55] K. Järvelin, J. Kekäläinen, Cumulated gain-based indicators of IR performance, Univ. Tampere, Dep. Inf. Stud. Res. Notes 2 (2002) 1–26. http://tampub.uta.fi/handle/10024/65718.

[56] J.L. Hintze, R.D. Nelson, Violin plots: a box plot-density trace synergism, Am. Stat. 52 (1998) 181–184, http://dx.doi.org/10.1080/00031305.1998.10480559.

[57] P. Sarchielli, I. Rainero, F. Coppola, C. Rossi, M. Mancini, L. Pinessi, et al., Involvement of corticotrophin-releasing factor and orexin-A in chronic migraine and medication-overuse headache: findings from cerebrospinal fluid, Cephalalgia 28 (2008) 714–722, http://dx.doi.org/10.1111/j.1468-2982.2008.01566.x.

[58] A.I. Scher, G.M. Terwindt, W.M.M. Verschuren, M.C. Kruit, H.J. Blom, H. Kowa, et al., Migraine and MTHFR C677T genotype in a population-based sample, Ann. Neurol. 59 (2006) 372–375, http://dx.doi.org/10.1002/ana.20755.

[59] N.C. Chai, B.L. Peterlin, A.H. Calhoun, Migraine and estrogen, Curr. Opin. Neurol. 27 (2014) 315–324, http://dx.doi.org/10.1097/WCO.0000000000000091.

[60] K.K. Hampton, A. Esack, R.C. Peatfield, P.J. Grant, Elevation of plasma vasopressin in spontaneous migraine, Cephalalgia. 11 (1991) 249–250. http://www.ncbi.nlm.nih.gov/pubmed/1790568.

[61] R.C. Peatfield, K.K. Hampton, P.J. Grant, Plasma vasopressin levels in induced migraine attacks, Cephalalgia 8 (1988) 55–57.

[62] D.S. Wishart, T. Jewison, A.C. Guo, M. Wilson, C. Knox, Y. Liu, et al., HMDB 3.0-the human metabolome database in 2013, Nucleic Acids Res. 41 (2013) 801–807, http://dx.doi.org/10.1093/nar/gks1065.

[63] C. Batchelor, C.Y.A. Brenninkmeijer, C. Chichester, M. Davies, D. Digles, I. Dunlop, et al., Scientific lenses to support multiple views over linked chemistry data, Semant. Web – ISWC 2014 8796 (2014) 98–113, http://dx.doi.org/10.1007/978-3-319-11964-9.

[64] M.J. Schuemie, M. Weeber, B.J.A. Schijvenaars, E.M. Van Mulligen, C. Van Der Eijk, R. Jelier, et al., Distribution of information in biomedical abstracts and full-text publications, Bioinformatics 20 (2004) 2597–2604, http://dx.doi.org/10.1093/bioinformatics/bth291.

[65] E. Cernuda-Morollón, D. Larrosa, C. Ramón, J. Vega, P. Martínez-Camblor, J. Pascual, Interictal increase of CGRP levels in peripheral blood as a biomarker for chronic migraine, Neurology 81 (2013) 1191–1196, http://dx.doi.org/10.1212/WNL.0b013e3182a6cb72.