Information from Semantic Integration of Texts and Databases

Erik M. van Mulligen[1], Wytze J. Vlietstra[1], Rein Vos[1,2], Jan A. Kors[1]

[1]Erasmus University Medical Center, Rotterdam, The Netherlands
{e.vanmulligen, j.kors, w.vlietstra}@erasmusmc.nl
[2]Maastricht University, The Netherlands
{rein.vos}@maastrichtuniversity.nl

Abstract. Relations mined from texts and structured information from databases have been mapped to concepts defined in biomedical ontologies and to a predicate dictionary. Concepts and predicates are represented by nodes and edges in this graph and can be queried for relations between concepts. The graph combines relations extracted from Medline abstracts with relations obtained from the UMLS and databases as UniProt, EntrezGene, Comparative Toxicogemics Database, and from the datasets from the Linked Open Drug Data (Drugbank, DailyMed, and Sider).

The approach was tested on 61 cerebral spinal fluid and 207 serum compounds of migraine patients. A cloud of all biomedical concepts related to the concept migraine in this graph was used to construct a set of cerebral spinal fluid compound concepts and a set of serum compound concepts. For each of the relations in the cloud provenance is available and provided. These sets were evaluated against two manually created sets of compounds.

The evaluation showed that this graph based method retrieves relevant compounds with mean average precision values of 0.32 and 0.59, respectively.

Keywords: graph databases, relation mining, Medline

1 Introduction

Are we not all dreaming about a computer program that, based on all available publications and data in databases, suggests the most likely hypotheses worth investigating in our application domain? And would it not be perfect if the program also provided us with an argumentation? In this paper we will outline the steps we have taken to support scientists in better understanding the information that is already available and how that could be used to generate new hypotheses.

The benefits and risks of the avalanche of information in the biomedical domain are widely recognized[1]. The potential value of all these data is that we are able to better understand the processes from the genetic level up to the disease or phenotype level.

The risk of having all these data available is that we lose sight of how the data relate and can be combined to provide new insights. An integrative approach is desirable on two levels: on the technological level by integrating numerous biological databases into networks of knowledge sources, and on the conceptual level by integrating different fields of biomedicine into networks of concepts. This integration can support new approaches to inference and search.

Swanson recognized the potential of relating disconnected fields of knowledge in biomedicine, in particular by discovering new associations between, as he called it, A and C terms, consisting of single words or short phrases (2-3 words). He developed a program, ArrowSmith[2], to automatically find B terms that co-occur with A and C terms in Medline titles. If the A and C terms were never co-mentioned in a title, a new potential discovery was identified. Using this approach he was able to discover a connection between Raynaud's disease (A) and fish oil (C) through blood coagulation (B), and between migraine (A) and magnesium (C) via blood clotting (B). These hypotheses were later on proven correct in experimental studies[3,4,5].

The value of this approach has been recognized by many scientists and a series of new research projects were started to improve on this. One method, explored by Blake and Pratt[6], was to use concepts as defined by the Unified Medical Language System (UMLS)[7], instead of separate terms. In the UMLS thesaurus, different terms that denote the same unit of thought have been normalized to a single concept. Weeber et al. were the first to mine concepts from both Medline titles as well as abstracts, by mapping terms to the UMLS thesaurus with the MetaMap concept recognizer[8]. Weeber et al. were also succesful in applying their system for a new discovery in drug research, suggesting thalidomide as a treatment for chronic hepatitis C, among others[9].

Swanson manually selected the B terms that he thought to be most relevant for further exploration. Many researchers have worked on approaches to automate the selection of the B terms. The concept-based approaches using UMLS have explored the use of the semantic types of the B concepts. Blake and Pratt used this approach to discard several semantic types and reported an 81% decrease of the number of B terms[10]. Srinivasan et al. applied a similar approach to filter out B terms based on semantic types[11]. If the relevant semantic types were precisely known, the set of terms could be reduced by as much as 91%; if only the obviously irrelevant semantic types were removed, the number of terms was reduced by an average of 31%. Gordon and Lindsay evaluated several ranking algorithms borrowed from the information retrieval field when they re-analyzed Swanson's *fish oil-Raynaud's Disease* discovery, such as Term Frequency-Inverse Document Frequency (TF-IDF)[12]. They reported reproduction of 10 of the 12 relevant B-terms for Swanson's discovery in a list of 35 terms.

To rank the B-terms, Torvik and Smalheiser applied an ensemble algorithm that combined eight weighted variables, such as "B-term occurs in more than one paper within literature sets A and C", "B-term maps to at least one UMLS semantic category", "B-term first appears recently within Medline as a whole", etc.[13] Swanson originally used a fixed order approach of first filtering uninformative terms using a stopword list, subsequently term categorization, and finally manual selection of B-terms. Instead Torvik developed this ensemble algorithm to containing all steps of Swanson's fixed order approach, having the advantage not to lose potentially relevant B terms in any of the intermediate steps.

Yetsigen-Yildiz et al. investigated different statistics to rank the B-terms[14]. Two of them were frequency-based association rules as tested by Hristovski[15] et al., and two were probability based, including the Z-score, which creates literature subsets, and the mutual information score. The association rules were not evaluated against the Swanson sets, but they were analyzed on their predictions from a subset of Medline future published discoveries.

Hristovski et al. were the first to test the added value of incorporating relation predicates into a literature-based discovery process[16]. They applied the UMLS semantic network and the SemRep text mining system to identify relationships between terms[17]. Predicates were used to identify discovery patterns: specific combinations of two predicates between three terms, which when combined would constitute a functional, biologically relevant association. Although the inclusion of predicates was considered to offer clear advantages, the lack of accuracy of the relationship extraction hampered practical application.

In our group we developed the Anni discovery system[18]. For each concept, the co-occurrences between that concept and other concepts in all Medline abstracts are computed and stored in a so-called concept profile. Concept profiles can be considered vectors in a high-dimensional vector space. The strength of the relationship between two concepts is expressed as a matching score between their concept profiles. Concepts can be grouped based on their semantic type and their concept profiles can be matched based on various algorithms: mutual information measure, log-likelihood, and dot product.[19] The matching strategy takes into account all the B concepts contained in the concept profile, filters the resulting C concepts on the required semantic type(s) and ranks the result on matching score. This approach has been used by Jelier et al. in a study to match the concept profiles for genes from DNA microarray data with concepts that denote functions of the genes[20]. The same approach has been used by Van Haagen et al. to predict protein-protein interactions by computing the matching score between protein concept profiles at certain time intervals in Medline[21]. An extension of this approach has been developed by using Anni in mapping disease-disease relationships for knowledge discovery in multi-morbidity research on somatic and psychiatric diseases[22].

The approach presented in this study is to combine relations obtained from literature with those available in databases and ontologies. The subject and object of each relation are mapped to a concept as defined in our ontology (mainly the UMLS with extensions for genes, proteins and chemicals). The predicates obtained from the text are mapped to a set of standardized predicates. The mapping process is partly supported by our text mining software and partly by manual mapping of database schemas to concepts and predicates.

2 Methods

Our approach combines relations extracted from Medline abstracts with relations obtained from the UMLS and databases UniProt, EntrezGene, Comparative Toxicogemics Database[23], and from the datasets from the Linked Open Drug Data[24] (Drugbank, DailyMed, and Sider) into a semantic graph database. From these different sources we identified 2,669,792 individual concepts, together with about 71 million relations between them. The relations are based on the 54 relationship types defined in the UMLS semantic network and the predicates defined by Halil and used in the Semantic Medline[25]. In total, 171 different predicates were defined. A concept consists of a set of terms (synonyms) that denote the concept, and identifiers that link to the various databases. Each concept is connected to one or more semantic type nodes in the graph database, a database that primarily consists of nodes and connections between nodes. Semantic types in turn are categorized in semantic groups[26].

The mapped relations are stored in our graph database. The graph database has been implemented in the Neo4J graph database, version 1.8.3[27]. We implemented a layer on top of Neo4J that implements the notion of concepts, labels, relations, semantic types, semantic groups, and provenance. Each relation − edge − between two concepts – nodes − has one or more of the semantic predicate labels and provenance information that indicates the source of the relation. Semantic predicates contain a direction and for both directions a set of labels is provided, typically the active and passive form of a verb. Neo4J has built in functionality to find paths between two nodes. We extended this functionality so that extra information − such as references to scientific articles that support the relation − can be included in evaluating the various paths.

2.1 Semantic Integration

We started building our graph database by incorporating the UMLS 2012AA (Metathesaurus and Semantic Network). We then proceeded by integrating Semantic Medline[28]. This source was easy to map to the UMLS concepts and to the semantic relations.

As an example of integrating a database we will outline how the mapping of UniProt to the graph database was done. A schematic representation of the database schema of UniProt is provided in Figure 1.

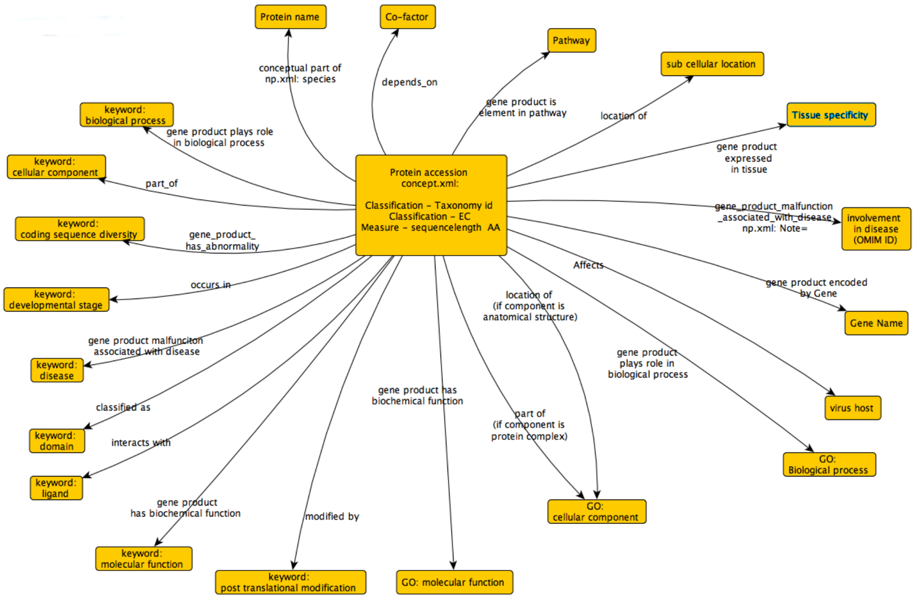

Fig. 1. Overview of the database schema of UniProt. The figure shows how the different aspects of the UniProt schema are mapped to a semantic relation and a UMLS concept.

The challenge of integrating UniProt entries lies in mapping the annotation fields to the corresponding UMLS concepts. We used our concept identification tool Peregrine[29] to find UMLS concepts in the free-text UniProt annotation fields. The mapping of the implicit relations defined in the UniProt schema to the proper semantic predicates is manual work and requires some understanding of the biological meaning of the data.

This mapping process has been repeated for all sources integrated thus far. We maintain a mapping database that indicates how identifiers from one coding system map to another coding system. These mappings make it easier to integrate a new resource if some of the fields are coded.

2.2 Inferencing

To use the graph database, we implemented a web service around it that provides basic functionality. In particular, for inferencing we implemented a path-finding algorithm based on Neo4J's functionality. This simple, path-finding type of inferencing is not following the main, logic-based inferencing approaches such as implemented with OWL-DL and formal reasoners. The extension of Neo4J's path-finding function allows one to specify a set of relations that restricts the set of edges that can be explored to find a path between the source and target concepts. The paths lengths are currently limited to a maximum of five edges. The path function can be modified and can take into account additional information that may influence the selection of edges, e.g., provenance information (the sources that support the relation).

In the remainder we will describe a knowledge discovery application that we developed. Experience with this application made clear that the inferencing should be tailored to the specific needs of the application domain. As mentioned by others[30] the user's semantic view is important for users of the graph database. The semantic view allows one to define the level of detail for particular groups of concepts. For example, a clinical researcher may not be so much interested in the fine-grained differences between a set of related chemical compounds but rather may want information at a higher abstraction level.

3 Results

We applied the graph database to a number of application domains. In this paper we selected the finding of new compounds marking the imminence of a migraine attack to demonstrate the use of the approach.

We obtained a set of 61 compounds that have been reported in the literature to be measurable in the cerebral spinal fluid, and a set of 207 compounds reported to be measurable in the serum of migraine patients. Both sets were manually constructed by a manual review process of a corpus of articles retrieved with PubMed., EMBASE and Web of Science. The objective was to test whether a graph database could be used to identify a set of linking concepts, similar to the linking B-terms, between these compounds and migraine. The question was whether this set of linking concepts with their interconnectivity could be used to identify (1) the original set of compounds, and (2) new compounds of interest. The two sets of compounds were fed to the graph database to obtain the paths between these compounds and migraine. These paths were analyzed for characteristics (number of publications, range of publication dates, path length, etc.). Additional compounds that were not part of the initial set have been viewed as potentially new discovered compounds.

The final result of this study was a set of concepts found in the paths linking migraine to these sets of compounds. A selection of this set of linking B-concepts was made on basis of the semantic types. Using this selected B-concept set we used the number of different connections between a compound and the B-concept set for reconstructing the initial given set of compounds and secondly to identify potential new compounds (see Figure 2). Several ranking statistics were evaluated and overall there was only very little difference. From the cerebral spinal fluid set of 61 compounds directly connected to mirgraine 1 could not be identified and from the serum set of 207 compounds directly connected to migraine 23 could not be identified using this approach. We computed a weighted mean average precision of 0.32 for the cerebral spinal fluid set and 0.59 for the serum set.

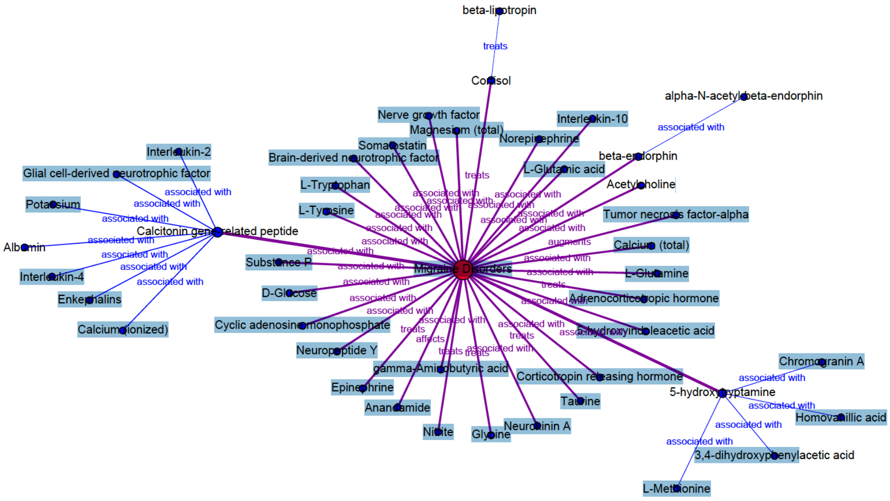

Fig. 2. Selection from the graph database showing the cloud of concepts linked to Migraine and the relations from this cloud to a number of compounds.

4. Discussion

As mentioned in the introduction when discussing the Swanson approach, the ranking and filtering of the B-terms determines to a large extent the success of the knowledge discovery method. A similar issue can be raised about the ranking and relevance of the connecting paths that our method constructs in a multi-source graph database. With increasing path lengths, at some point each pair of concepts in the graph database will be connected. It will therefore be important to investigate approaches that can differentiate between useful and sound discovery paths and those that are noisy and redundant. The platform is powerful in its potential to implement discovery patterns that combine a rich feature set consisting of semantic types, semantic groups, semantic predicates, connectivity, and amount of provenance stemming from different sources.

From our experiments thus far it became clear that a more formal framework to the relations or semantic predicates would be helpful. Similar to semantic types and groups, which denote the specific properties of concepts, we may imagine that logic classes on top of the predicates would indicate specific properties of the predicates, such as transitiveness. A framework that follows a more logic-based foundation is the OpenBEL framework[31]. In future work we will assess whether our semantic predicates can be mapped to this framework.

For this application we did not restrict the discovery connection paths on basis of the combination of a particular semantic groups or types of concepts with a set of particular predicates. Our first experience is that such a selection might help in finding more relevant connections. The flexibility of the graph database to support various types of selections has been used in an application in the field of adverse drug reactions and in food safety. We will further investigate in how far these selections are depending on an application and can be formalized in a guideline on how to use a graph database for discovery.

5. Conclusions

The graph database that we constructed combines information extracted from biomedical texts with information obtained from biological databases. We have demonstrated in this paper that relations from texts and structured databases can be effectively combined in a single graph database. Our inferencing approach illustrated in this paper shows that relevant compounds can be retrieved with a fairly high recall. Furthermore, our approach shows that the connectivity to a set of other concepts has potential. The flexibility of the graph database makes it possible to apply the approach to other discovery applications and evaluate other approaches to combine graph statistics and filters on semantic groups and predicates..

References

1. Lu Z. Pubmed and beyond: a survey of web tools for searching biomedical literature. Database 2011, Oxford.
2. Swanson DR, Smalheiser NR. An interactive system for finding complementary literatures: a stimulus to scientific discovery. Artificial Intelligence, 1997;91:183-203.
3. Swanson DR. Fish oil, Raynaud's syndrome, and undiscovered public knowledge. Perspect Biol Med. 1986;30(1):7-18.
4. Swanson DR. Migraine and magnesium: eleven neglected connections. Perspect Biol Med. 1988; 31(4):526-557.
5. Swanson DR. Medical literature as a potential source of new knowledge. Bull Med Libr Assoc. 1990;78(1):29-37
6. Blake C, Pratt W. Automatically identifying candidate treaments from existing medical literature. AAAI Spring Symposium on Mining Answers from Texts and Knowledge Bases. 2002;9-13.
7. Bodenreider O. The unified medical language system (UMLS): Integrating biomedical terminology. Nucleic Acids Res. 2004;32(Database issue):D267-D270
8. Weeber M, Klein H, de Jong-van den Berg LTW, Vos R. Using concepts in literature-based discovery: Simulating Swansons Raynaud-Fish Oil and Migraine-Magnesium Discoveries. *J Am Soc Inf Sci Technol.* 2001;52(7):548-557
9. Weeber M, Vos R, Klein H, Aronson AR, Molema G. Generating hypotheses by discovering implicit associations in the literature: a case report of a search for new potential therapeutic uses for thalidomide. Journal of the American Medical Informatics Association 2003;10(3):252-259.

10. Blake C, Pratt W. Automatically identifying candidate treatments from existing medical literature. AAAI Spring Symposium on Mining Answers from Texts and Knowledge Bases. 2002;9-13.
11. Srinivasan P. Text mining: generating hypotheses from Medline. Journal of the American Society for Information Science and Technology. 2004;55(5):396-413.
12. Lindsay RK, Gordon MD. Literature-based discovery by lexical statistics. 1999.
13. Torvik VI, Smalheiser NR. A quantitative model for linking two disparate sets of articles in Medline. Bioinformatics. 2007;23(13):1658-1665.
14. Yetisgen-Yildiz M, Pratt W. A new evaluation methodology for literature based discovery systems. J Biomed Inform. 2009;42(4):633-643.
15. Hristovski D, Stae J, Peterlin B, Dzeroski S. Supporting Discovery in Medicine by Association Rule Mining in Medline and UMLS. Medinfo 2003;10(Pt2):1344-1348.
16. Hristovski D, Friedman C, Rindflesch TC, Peterlin B. Exploiting semantic relations for literature-based discovery. AMIA annual symposium proceedings. 2006;349.
17. Ahlers CB, Fiszman M, Demner-Fushman D, Lang F, Rindflesh TC. Extracting semantic predication from MEDLINE citations for pharmacogenomics. In Pacific Symposium on Biocomputing. 2007:209–220.
18. Jelier R, Schuemie MJ, Veldhoven A, Dorssers LC, Jenster G, Kors JA. Anni 2.0: a multipurpose text-mining tool for the life sciences. Genome Biol. 2008;9(6):R96.
19. Jelier R, Schuemie MJ, Roes PJ, van Mulligen EM, Kors JA. Literature-based concept profiles for gene annotation: The issue of weighting. Int J of Med Inform 2008;77(5):354–362.
20. Jelier R, Jenster G, Dorssers L, Wouter B, Hendriksen P, Mons B, Delwel R, Kors J. Text-derived concept profiles support assessment of DNA microarray data for acute myeloid leukemia and for androgen receptor stimulation. BMC Bioinformatics. 2007;8:14.
21. van Haagen HH, 't Hoen PA, de Morrée A, van Roon-Mom WM, Peters DJ, Roos M, Mons B, van Ommen GJ, Schuemie MJ. In silico discovery and experimental validation of new protein-protein interactions. Proteomics. 2011;11(5):843-53.
22. Vos R, Aarts S, van Mulligen EM, Metsemakers J, van Boxtel MP, Verhey F, van den Akker MJ. Finding potentially new multimorbidity patterns of psychiatric and somatic diseases: exploring the use of literature-based discovery in primary care research. J Am Med Inform Assoc. 2014; 21(1):139-45.
23. Davis AP, Murphy CG, Johnson R, Lay JM, Lennon-Hopkins K, Sara-ceni-Richards C, Sciaky D, King BL, Rosenstein MC, Wiegers TC, Mattingly CJ. The Comparative Toxicogenomics Database: update 2013. Nucleic Acids Res. 2013;41(D1):D1104-14.
24. http://www.w3.org/wiki/HCLSIG/LODD
25. Kilicoglu H, Rosemblat G, Fiszman M, Rindflesch TC. Constructing a semantic predication gold standard from the biomedical literature. BMC Bioinformatics. 2011;12:486.
26. McCray AT, Burgun A, Bodenreider O. Aggregating UMLS semantic types for reducing conceptual complexity. Stud Health Technol Inform. 2001;84(Pt 1):216-20.

27. Neo4J Developers: Neo4J, Graph NoSQL Database 2012. Available at: http://neo4j.org/.
28. Cairelli MJ, Miller CM, Fiszman M, Workman TE, Rindflesch TC. Semantic MEDLINE for discovery browsing: using semantic predications and the literature-based discovery paradigm to elucidate a mechanism for the obesity paradox. AMIA Annu Symp Proc. 2013;164-73.
29. https://trac.nbic.nl/data-mining/
30. Brenninkmeijer CYA, Evelo C, Goble C, Gray AJG, Groth P, Pettifer S, Stevens R, Williams A, Willighagen EL. Scientific Lenses over Linked Data: An approach to support task specific views of the data. A vision. In: Proceedings of the 2nd International Workshop on Linked Science 2012 – Tackling Big Data (LISC2012), in conjunction with 11th International Semantic Web Conference (ISWC2012). 2012, Boston, MA.
31. http://www.openbel.org