Towards using a Graph Database and Literature-based Discovery for Interpretation of Next Generation Sequencing Results

Dimitar Hristovski
Medical faculty
Ljubljana, Slovenia
dimitar.hristovski@mf.uni-lj.si

Gaber Bergant
KIMG, UMC Ljubljana
Ljubljana, Slovenia
gaber.bergant@kclj.si

Andrej Kastrin
Medical faculty
Ljubljana, Slovenia
andrej.kastrin@guest.arnes.si

Borut Peterlin
KIMG, UMC Ljubljana
Ljubljana, Slovenia
borut.peterlin@kclj.si

***Abstract*—The arrival of high-throughput sequencing technologies in routine diagnostic medicine has enabled the large scale use of these technologies; however, the challenges of interpreting the results for diagnostic purposes has recently become evident. For this reason, we aim to develop a bioinformatics tool for clinical genetics diagnostics support. We gathered the data for this project from several different sources, including semantic relations extracted with SemRep from the MEDLINE bibliographic database, clinical phenotype observations from clinical geneticists and finally genotype data produced by Next-Generation Sequencing (NGS) annotated with population data and several theoretical pathogenicity prediction algorithms. We stored this data in a Neo4j graph database and employed it using a closed discovery approach to Literature-Based Discovery (LBD) as a complementary method in diagnostic NGS data analysis. All algorithms were implemented using the Cypher query language. The goal of the study was first to determine the usability of graph databases to represent heterogeneous clinical genomic data and secondly to determine the potential benefits of using LBD as a complementary approach in a diagnostic setting using NGS.**

Keywords- Literature-based discovery; Next-generation sequencing; Genomic data analysis; Semantic MEDLINE; Graph database; Neo4j.

I. Introduction

Next Generation Sequencing (NGS), also known as high-throughput sequencing, is a term collectively describing several different technologies that integrate a massively parallel sequencing approach, thus enabling the sequencing of whole human genomes within reasonable timescales. The development of NGS technologies successfully spread the utility of (clinical) Deoxyribonucleic Acid (DNA) sequencing by reaching unprecedented speed at reduced cost, enabling wide spread clinical and research use and thus fueling the rapid growth of genomic sciences. This presents a challenge in that pursuing and incorporating the newly discovered data in analytical pipelines becomes more and more difficult. Targeting this issue, we present our preliminary research in using the Literature-Based Discovery (LBD) paradigm to improve the interpretation of NGS results.

II. Methods

The goal of LBD is to generate novel hypotheses by analyzing the literature and optionally other knowledge sources [1]. For a recent review of LBD tools and approaches see [2]. We can approach LBD with one of two paradigms, either open or closed discovery. For this project we selected closed discovery. We also chose Neo4j [3] as a graph database for storing the collected data. Briefly, we deal with the data from a single patient at a time. The input is two sets of data for each patient, the genotype of discovered genomic variants and the phenotype as observed by the clinical geneticist. The genotype set X contains the genes with mutations as found by diagnostic NGS. The phenotype set Z contains the clinical observations provided by the clinical geneticist described using the human phenotype ontology (HPO) terms.

After gathering the relevant datasets, we constructed a graph database in Neo4j. The graph database consists of two major types of nodes, patients and concepts of several types including phenotypes, genes, proteins, cell functions, genetic disorders and many other biomedical types. Connecting these nodes, we have several different relationship types. For example, the relationship PHENO connects patients with their corresponding phenotype nodes and the relationship GENO connects patients with their respective mutated genes. This relationship also holds the information of the variant location, the specific mutated nucleotide, the severity of the mutation type (mainly differentiating missense and loss of function variants), predictions of theoretical algorithms for pathogenicity prediction and population data from the repository of the Genome Aggregation Database (gnomAD) [4]. We used this additional information for filtration and prioritization of candidate genes, thus reducing the workload required for manual result review by a clinical expert. Additionally, we have included the 30 different types of semantic relations as extracted by SemRep [5] from all of MEDLINE (titles and abstracts) serving as a backbone for patient and phenotype node connection. SemRep is a rule-based, symbolic natural language processing system that extracts semantic relationships in clinical medicine, substance interactions, genetic etiology of disease, and pharmacogenomics (e.g., TREATS, INHIBITS, STIMULATES, CAUSES, PREDISPOSES, AUGMENTS). These relationships are publicly available as SemMedDB [6] (a MySQL database). This work is a continuation and extension of our previous work [7] in which we explained how to construct a Neo4j graph database from SemMedDB and how to implement generic LBD with Cypher.

We assumed that the feasibility of our approach in routine clinical practice depended heavily on the development of intuitive algorithms for final result filtration and prioritization. We also expected this to prove especially useful in the clinical setting where expert time is limited. Therefore, we implemented a practical prioritization algorithm, which outputs an ordered list of genes awaiting expert review. For this purpose, we used the integrated relevant statistical data, such as population frequencies extracted from several population databases encompassing worldwide healthy control populations as well as theoretical pathogenicity predictions from several available prediction algorithms. Finally, we programmatically employed the prioritization algorithm in the Cypher query language on a per patient basis, extracting a prioritized list of genes, hopefully leading to further diagnostic and research steps, possibly also improving the diagnostic yield of NGS.

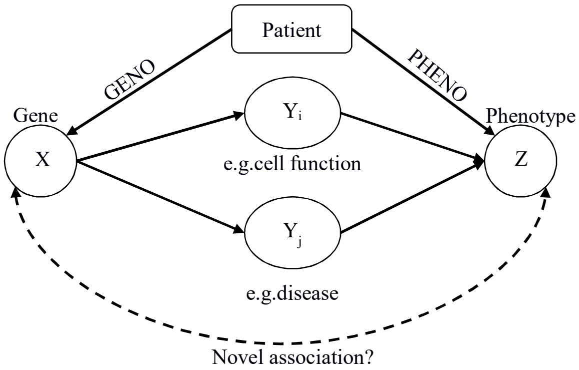

Figure 1. Illustration of novel clinical association prediction. Based on clinical genetic data and current biomedical knowledge (solid lines) we predict novel clinical associations (dashed arcs).

Figure 1 illustrates our approach. The output of the algorithm is a set of relevant intermediate concepts Y (such as genetic functions and/or diseases) that link the genotype X to the phenotype Z. These Y concepts should provide a hypothesis that explains the mechanisms for the novel associations that link the genotype to the phenotype. Our algorithm is meant as a discovery support step in a clinical NGS data processing pipeline. We generate hypotheses (explanations and novel associations); however, a knowledgeable human expert is needed for the critical evaluation of these hypotheses.

III. Results

The network we constructed consists of 1205 patient nodes. There are 262132 GENO relationships linking the patients to 15294 gene nodes. Multiple GENO relations are possible between a particular patient and a particular gene because sometimes there are multiple mutations (genetic variants) in the same gene for a particular patient. There are 4751 PHENO relations between the patients and the corresponding phenotypes represented with 1450 HPO terms (represented as nodes). The 1450 HPO terms were mapped to 982 UMLS concepts, which are used as arguments in the SemRep semantic relations. With SemRep we extracted 91567597 semantic relation instances from 55551193 sentences from 27263265 MEDLINE bibliographic records. The semantic relations instances were aggregated into 20818782 semantic relations between 277160 biomedical concept nodes, and were stored in our Neo4j database. The SemRep semantic relations represent general biomedical knowledge. The clinical patient data (the genotype and phenotype) are linked to the general biomedical knowledge through common concepts (represented as nodes).

Essential for the clinical implementation of the graph database was the storage of genotype data within multiple relationships between the patient and the specific gene, where the mutations occurred, as that allowed for the possibility of storing sequencing information regardless of the number of variants occurring in this gene.

IV. Discussion and further work

In this preliminary work, we constructed the necessary bioinformatics infrastructure in such a way that it allows efficient import of clinical data and simple extraction of relevant results, which are necessary for its inclusion in the diagnostic pipeline. However, several challenges remain to be solved. From the clinical genomics perspective, we noticed that within the SemMed database, the relationships between some genes and their correlated biomedical concepts are underrepresented while other genes are highly connected with non-informative concepts. Another issue is that some concepts are too general and highly connected to other concepts rendering them non-informative. We plan to find ways for filtering out non-informative concepts by using network analysis centrality measures and community detection algorithms. Furthermore, within the ranking algorithm, we face Cypher language efficiency issues, which we plan to address by query optimization. We also plan to develop an interactive visualization tool, enabling a quick and easy visual analytics. And finally, we plan to evaluate the usefulness of our approach from the clinical genetics point of view.

V. Conclusion

Although the LBD paradigm has been used in a research context for some time, it has been underappreciated in the clinical genetic diagnostic setting. With this preliminary study we show the potential of using LBD as a complementary method in clinical diagnostics of genetic disorders, with the emphasis on novel gene-phenotype associations. Furthermore, we determined that using a graph database such as Neo4j is suitable for storing heterogeneous genomic data needed for clinical genetics diagnostic support.

Acknowledgment

We thank Thomas C. Rindflesch for his comments and suggestions regarding this manuscript.

References

[1] D. R. Swanson, “Fish oil, Raynaud’s syndrome, and undiscovered public knowledge,” Perspectives in Biology and Medicine, vol. 30, no. 1, pp. 7-18, 1986.

[2] D. Hristovski, T. Rindflesch, and B. Peterlin, “Using literaturebased discovery to identify novel therapeutic

approaches," Cardiovascular & Hematological Agents in Medicinal Chemistry, vol. 11, no. 1, pp. 14-24, 2013.

[3] Neo4j website. Available at: http://neo4j.com. Last accessed March 10th 2018.

[4] M. Lek, K. J. Karczewski, E. V. Minikel et al. Exome Aggregation Consortium. Analysis of protein-coding genetic variation in 60,706 humans. Nature. 2016 Aug 18;536(7616):285-91.

[5] T. C. Rindflesch and M. Fiszman, "The interaction of domain knowledge and linguistic structure in natural language processing: Interpreting hypernymic propositions in biomedical text," Journal of Biomedical Informatics, vol. 36, no. 6, pp. 462-477, 2003.

[6] H. Kilicoglu, D. Shin, M. Fiszman, G. Rosemblat, and T. C. Rindflesch, "SemMedDB: A PubMed-scale repository of biomedical semantic predications," Bioinformatics, vol. 28, no. 23, pp. 3158-3160, 2012.

[7] D. Hristovski, A. Kastrin, D. Dinevski, and T. C. Rindflesch, "Towards implementing semantic literature-based discovery with a graph database," Proceedings of the GraphSM 2015, The Second International Workshop on Large-scale Graph Storage and Management, pp. 180-184, 2015.