Using Predicate Information from a Knowledge Graph to Identify Disease Trajectories

Vlietstra, W.J.[1], Vos, R.[1,2], van Mulligen, E.M.[1], Kors, J.A.[1]

[1] Department of Medical Informatics, Erasmus Medical Centre, Rotterdam, 3015 GE, the Netherlands

[2] Department of Methodology & Statistics, Maastricht University, Maastricht, 6229 HA, the Netherlands

w.vlietstra@erasmusmc.nl

Introduction

Knowledge graphs can represent the contents of biomedical literature and databases as subject-predicate-object triples, where predicates describe relationships between pairs of biomedical entities. For example, the Reactome database contains the triple "GTF2H2-controls the expression of-MDC1", and SemMedDB, which obtains its triples through text-mining, contains the triple "IL1B-stimulates-MCP1". By integrating triples from different sources with each other in a knowledge graph, the comprehensive body of biomedical knowledge can be computationally analyzed.

Analyses performed on knowledge graphs often aim to identify new relationships, e.g., between drugs and diseases, genes and phenotypes, or between diseases. However, from large-scale observational studies we know that multiple diseases in patients are often diagnosed in specific temporal sequences, which are referred to as disease trajectories. Using knowledge graphs to identify disease trajectories therefore requires both identifying the correct pair of diseases, as well as their correct temporal sequence.

Because protein networks are involved with metabolic, signaling, immune, and gene-regulatory networks, they are often used to mechanistically explain relationships between diseases. So-called disease proteins, which are proteins coded by genes associated with a disease, can be used to represent diseases on a protein level. However, until now predicates between proteins have been rarely used, even though they, by describing the relationships between (disease) proteins, can provide additional information about the mechanism by which one disease can lead to another. We therefore aim to exploit the predicate information from paths between (disease) proteins in a knowledge graph to determine whether a sequence of two diseases forms a trajectory.

Method

The temporal disease trajectories as described by Jensen et al. were used as a reference set (Jensen 2014). They analyzed diagnoses in 6.2 million electronic patient records of the Danish population, assigned during 14.9 years, to identify common disease trajectories. From these trajectories, we only used those that describe a sequence of two diseases. A complementary, negative set of non-trajectories was constructed by creating random pairs of the diseases in the reference set, as well as the reversed (incorrect) temporal sequence of the trajectories in the reference set. Associations between proteins and diseases were obtained from the manually curated subset of DisGeNet (Piñero 2017).

Three scenarios of paths between the disease proteins of pairs of diseases were extracted from the knowledge graph:

1) Overlap, where two diseases A and B share the same disease protein. Optionally, this disease protein has a relationship to itself, e.g., if it can homodimerize.
2) Direct path, where there is a triple of which one of the disease proteins of disease A and one of the disease proteins of disease B form the subject and object.
3) Indirect path, where one intermediate protein connects the disease proteins of disease A and disease B, requiring a sequence of two triples.

Based on the predicates within these paths, six feature sets were constructed. We compared two methods to represent indirect relationships between disease proteins. The first method constructs so-called metapaths (Himmelstein 2017), where the sequence of predicates in an indirect path is used as a single feature. The second method considers each predicate in the indirect paths as a separate feature (Vlietstra 2018).

Table 1 Classification results for the six feature sets when trained with balanced training sets. The values in the AUC columns indicate mean ROC-area under the curve values of 10 repeats of a 10-fold cross validation experiment, along with their standard deviation.

	Metapaths		Split paths	
	Number of features	AUC	Number of features	AUC
Undirected	1217	81.3% (1.4%)	168	76.0% (1.5%)
Mixed	2823	87.9% (0.9%)	257	84.0% (1.1%)
Directed	3773	88.1% (0.9%)	277	81.7% (1.5%)

For both methods we experimented with three variations of directional information of the predicates. Directional information was never used when the same protein was both subject and object of the triple (overlap scenario).

1) Undirected: triples forming direct and indirect paths between disease proteins are used without information about which proteins are subject and object.
2) Directed: Each triple, in each direct and indirect path between the disease proteins, has a direction as indicated by its subject and object.
3) Mixed: Each predicate in the direct and indirect paths is classified as directed or undirected based on prior information.

Results

Our reference set contained 2,530 trajectories and 168,870 non-trajectories. We used random forests to train a classification model. The cross-validated performance is shown in Table 1, along with the number of features in the feature set. Use of directional information of predicates substantially improved performance as compared to not using this information. However, disease trajectories could still be identified with reasonable performance if only undirected information was used.

The metapath feature sets consisted of 7 to 14 times more features than the split-path feature sets, and achieved a superior performance as compared to the split-path features. The performance difference between the mixed and the directed metapath features was negligible. The performance of split features increased if prior knowledge about directed or undirected predicates was taken into account.

Discussion

Our work demonstrates that disease trajectories can be identified using predicate information from a protein knowledge graph. Our machine learning based classifier is capable of both identifying the correct pairs of diseases, as well as their correct temporal sequence. While the use of directional information of triples in our analysis improved performance, even when no directional information is used our classifier can identify directed relationships with reasonable performance. The use of prior knowledge to classify predicates as directed or undirected improves performance on split path feature sets, but has no impact with metapath feature sets. Metapaths result in many more features than the split paths, and consistently achieve a superior performance.

As future work we intend to perform a detailed error analysis, where we will investigate whether there are specific diseases whose trajectories are frequently misclassified. The International Classification of Diseases (ICD) hierarchy can be used to abstract diseases to a higher ICD level, thereby obtaining insight into misclassifications at the level of disease classes. Abstracting the diseases in the trajectories also allows to examine whether specific combinations of ICD classes are more frequently misclassified.

References

Himmelstein, D.S., Lizee, A., Hessler, C., Brueggeman, L., Chen, S.L., Hadley, D., Green, A., Khankhanian, P., Baranzini, S.E. 2017. Systematic integration of biomedical knowledge prioritizes drugs for repurposing. In *eLIFE*, 6:1-35

Jensen, A.B., Moseley, P.L., Oprea, T.I., Ellesøe, S.G., Eriksson, R., Schmock, H., Jensen, P.B., Jensen, L.J., Brunak, S. 2014. Temporal disease trajectories condensed from population-wide registry data covering 6.2 million patients. In *Nature Communications*, 5:4022

Piñero, J., Bravo, À., Queralt-Rosinach, N., Gutiérrez-Sacristán, A., Deu-Pons, J., Centeno, E., García-García, J., Sanz, F., Furlong, L.I. 2017. DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants. In *Nucleic Acids Research*, 45:833-839

Vlietstra, W.J., Vos, R., Sijbers, A.M., van Mulligen, E.M., Kors, J.A. 2018. Using predicate and provenance information from a knowledge graph for drug efficacy screening. In *Journal of Biomedical Semantics*, 9:23