As Ontologies Reach Maturity, Artificial Intelligence Starts Being Fully Efficient: Findings from the Section on Knowledge Representation and Management for the Yearbook 2018

Ferdinand Dhombres[1,2], Jean Charlet[1,3], Section Editors for the IMIA Yearbook Section on Knowledge Representation and Management

[1] Sorbonne Université, Université Paris 13, Sorbonne Paris Cité, INSERM, UMR_S 1142, LIMICS, Paris, France
[2] Sorbonne Université Médecine, Service de Médecine Fœtale, AP-HP/HUEP, Hôpital Armand Trousseau, Paris, France
[3] AP-HP, DRCI, Paris, France

Summary

Objectives: To select, present, and summarize the best papers published in 2017 in the field of Knowledge Representation and Management (KRM).

Methods: A comprehensive and standardized review of the medical informatics literature was performed to select the most interesting papers of KRM published in 2017, based on a PubMed query.

Results: In direct line with the research on data integration presented in the KRM section of the 2017 edition of the International Medical Informatics Association (IMIA) Yearbook, the five best papers for 2018 demonstrate even further the added-value of ontology-based integration approaches for phenotype-genotype association mining. Additionally, among the 15 preselected papers, two aspects of KRM are in the spotlight: the design of knowledge bases and new challenges in using ontologies.

Conclusions: Ontologies are demonstrating their maturity to integrate medical data and begin to support clinical practices. New challenges have emerged: the query on distributed semantically annotated datasets, the efficiency of semantic annotation processes, the semantic representation of large textual datasets, the control of biases associated with semantic annotations, and the computation of Bayesian indicators on data annotated with ontologies.

Keywords

Knowledge representation (computer); biomedical ontologies; controlled vocabularies; decision support systems, clinical; information storage and retrieval; data integration

Yearb Med Inform 2018:140-5
http://dx.doi.org/10.1055/s-0038-1667078

Introduction

The year 2017 has produced a large amount of publications related to Knowledge Representation and Management (KRM) in medicine. KRM focuses on the development of techniques to be used and leveraged in other medical informatics domains. In recent years, especially the last two years, we have observed an increasing number of works combining ontology engineering and supervised learning technologies. In this context, the nature and impact of ontology is discussed in different papers.

In this review, we present a selection of some of the best papers published in 2017 in the KRM domain, based either on their impact or the novelty of the approach proposed in the medical knowledge representation and management field.

Paper Selection Method

We conducted the selection of KRM papers in PubMed/MELDINE based on the same query as in the previous edition of the International Medical Informatics Association (IMIA) Yearbook [1]. We followed a generic method, commonly used in all sections of the Yearbook since 2013. As for the last four years, the search was performed on MEDLINE by querying PubMed. Our query includes MeSH descriptors related to the KRM field in the context of medical informatics with a restriction to international peer-reviewed papers, including international conference proceedings indexed in PubMed. Only original research articles published in 2017 (from 01/01/2017 to 12/31/2017) were considered; we excluded the following publications types: reviews, editorials, comments, and letters to the editor. We limited the search on the major MeSH descriptors (for example "biomedical ontologies [MAJR]") to avoid retrieving a large set of articles, and we completed it by non-MeSH terms searched on the titles and abstracts of the articles (for example "terminologies [TIAB]").

The selection of best papers was performed in a three-step process among the papers retrieved by the query. At the first step, section editors reviewed all titles, abstracts, and types of publications to establish a short list of 15 candidate best papers. At the second step, five experts (including the two section editors) reviewed the candidate best papers using the IMIA Yearbook quality criteria scoring method. More specifically, the following aspects of papers were evaluated: significance, quality of scientific content, originality and innovativeness, coverage of the related literature, organization and quality of the presentation. The final step of the selection was achieved during a meeting of the whole editorial board, based on the external reviews and the report of the two section editors.

Results

The KRM query retrieved 1,998 citations from PubMed. It accounts for a 41% increase in comparison with the results of the same query applied on papers published in 2016. Section editors achieved a first selection of 100 papers based on titles and abstracts. After a second review of this set of papers, including full text reviews, a selection of 15 candidate best papers was established [2-16]. Five reviewers scored these 15 pre-selected papers and finally selected five final best papers [2-6].

In direct line with the research on data integration presented last year [1], the 2017's best papers demonstrate even further the added-value of ontology-based integration approaches for phenotype-genotype association mining. Three papers describe systems [2, 3, 6]. The first paper presents a system that exploits semantic technologies with automated reasoning over genotype-phenotype relations to prioritize variants in whole exome and whole genome sequencing datasets [2]. The second paper describes a system which predicts the associations between human genes and phenotypes in Human Phenotype Ontology (HPO) based on human protein–protein interaction network [6]. The third paper presents a system which assembles high-throughput sequencing and microarray data by measuring the semantic similarity between different samples, with the aim of combining experiments associated with similar semantic annotations [3]. The fourth paper introduces a novel method to capture the hierarchical relationships between ontology classes during the learning process, thus allowing novel associations between genes and abnormal phenotypes prediction, with a significant reduction of computational costs [5]. In the fifth paper, the authors introduce a query federation engine that enables policy-aware access to healthcare sensitive data sets represented as Resource Description Framework (RDF) data cubes [4].

Among the ten other selected papers for 2017, we observed two other directions in the research conducted in the KRM field, in addition to the research on ontology-based data integration for phenotype-genotype association mining. One is the design of knowledge bases and their application, a "traditional" topic in our field. The other describes a series of emerging challenges associated with the growing use of ontologies.

In the next paragraphs, we grouped the selected papers in these three dimensions of the KRM research: the mining of genotype-phenotype associations, the design of knowledge bases, and the new challenges in using ontologies.

Table 1 Best paper selection of articles for the IMIA Yearbook of Medical Informatics 2018 in the section 'Knowledge Representation and Management'. The articles are listed in alphabetical order of the first author's surname.

Section
Knowledge Representation and Management
▪ Boudellioua I, Mahamad Razali RB, Kulmanov M, Hashish Y, Bajic VB, Goncalves-Serra E, Schoenmakers N, Gkoutos GV, Schofield PN, Hoehndorf R. Semantic prioritization of novel causative genomic variants. PLoS Comput Biol 2017;13(4):e1005500.
▪ Galeota E, Pelizzola M. Ontology-based annotations and semantic relations in large-scale (epi)genomics data. Brief Bioinform 2017;18(3):403-12.
▪ Khan Y, Saleem M, Mehdi M, Hogan A, Mehmood Q, Rebholz-Schuhmann D, Sahay R. SAFE: SPARQL Federation over RDF Data Cubes with Access Control. J Biomed Semantics 2017;8(1):5.
▪ Notaro M, Schubach M, Robinson PN, Valentini G. Prediction of Human Phenotype Ontology terms by means of hierarchical ensemble methods. BMC Bioinformatics 2017;18(1):449.
▪ Petegrosso R, Park S, Hwang TH, Kuang R. Transfer learning across ontologies for phenome-genome association prediction. Bioinformatics 2017;33(4):529-36.

Ontology-based Integration Approaches for Phenotype-genotype Association Mining

Four of the five best papers are presenting research in this specific domain [2, 3, 5, 6]. The papers are summarized in the Appendix of this synopsis. The use of HPO was the reference for the phenotyping in three papers. In the paper from Galeota and Pelizzola [3], the use of Open Biological Ontologies (OBO) ontologies-based tools showed better results than Unified Medical Language System (UMLS)-based tools for the phenotype annotations. This might be a consequence of using UMLS version 2014AA released before the HPO integration to UMLS (in version 2015AB and later) and also a consequence of using topic-specific OBO ontologies.

The paper of Boudellioua et al. [2] was the best rated of the selection. It should be noted that it is the one that built the most complex models by using several ontologies and by using a rather complex process of integration of those ontologies in a formally valid final file. The originality of the paper of Galeota and Pelizzola [3] lies in the comparison of the annotation resources used. If the comparison between UMLS and OBO ontologies can be discussed, it allows to question the viability of resources developed specifically versus initiatives, like UMLS, which seek to integrate widely specific resources with long term maintenance. From this point of view, the previous article [2] has chosen to make itself the integration of all needed resources. It is important to note that the article by Notaro et al. [5] and the one of Petegrosso et al. [6] relate to very similar subjects that lead the authors to develop learning algorithms that take into account the hierarchical structure of ontologies, be it HPO or Gene Ontology (GO). Overall, these four papers confirm a fair maturity in available resources for the operational semantic description of phenotypes.

In the same direction, the paper from Alonso-Calvo *et al.* [7] addresses the integration of genomic and clinical data in European data centers, by developing a semantic interoperability layer which uses standard terminologies and standards for information management (*e.g.* Health Language Seven (HL7), Relational Data Bases to RDF Modeling Language (R2RML)).

Ontologies in Action: Knowledge Base Design and Applications

Zhang *et al.* [16] introduced an ontology-based framework to integrate patient data, medical domain knowledge, and patient assessment criteria for chronic disease patient follow-up assessments. This framework was instantiated using real clinical data (115K follow-up assessment records of 36K type 2 diabetic patients) and resulted in a clinical decision support system (CDSS) for the automatic selection and adaptation of standard assessment protocols to suit patient personal conditions. The system demonstrated significant performances (accuracy of 99.93%, completeness of 95.00%), thus contributing to the improvement of the accessibility, the efficiency, and the quality of patient follow-up services. This approach being generic to knowledge sharing and reuse for patient-centered chronic disease management, this work paves the way for the development of CDSS for the care of other chronic disease patients. In the same vein of leveraging semantic representation to support clinical practices, Esteban-Gil *et al.* [10] presented a semantic interoperability platform with a real focus on usability for clinicians. The platform design is detailed with relevant choices of Semantic Web techniques supporting interactively manipulatable decision support (and visualizations) derived from cancer registries data. As of today, this promising platform was assessed only on simulation data and awaits for a clinical evaluation. These two systems demonstrated the efficacy of using semantic representations of data to develop CDSS. In 2017 however, it appears that these systems, some of them being fully functional, are still at early stages of deployment, with poor usability in clinical practice.

In contrast, the development of knowledge bases and ontological resources has become usual in the KRM medical community. Four years after the initial publication of the Protein Ontology (PRO), Natale *et al.* [13] presented the recent developments of this OBO resource. This group developed a standardized description of proteoforms, based on a specific syntax, that enforces the interoperability with Reactome (a biological pathway reference resource) and consequently with Reactome-related resources (Open Targets, Chemical Entities of Biological Interest (ChEBI) and UniProt). This significant evolution enhanced the coverage of PRO up to ~60% of Reactome proteoforms. Interoperability with other resources was also implemented via new pipelines, including dynamic generation of terms. Among many evolutions, the PRO team has addressed community needs and provided an Ontology Web Language (OWL) version and a SPARQL Protocol and RDF Query Language (SPARQL) endpoint. The Immune Epitope Database (IEDB) project is another example of knowledge base. It was presented by Vita *et al.* [15], illustrating the practical consequences of good practices in controlled vocabularies integration, with curation processes simplification and efficient interoperability between the IEDB and other resources. Similarly, Gipson and his group [11] developed a terminology for pediatric adverse events (PAEs). Although without significant novelty from the perspective of research in terminology design, this paper emerges as a good illustration of an international collaboration for achieving a shared PAE terminology with an appropriate integration within Medical Dictionary for Regulatory Activities (MedDRA) and other UMLS terminologies.

New Challenges in Using Ontologies

With the large adoption of ontologies in the KRM process, new challenges both technical and methodological arose. Two technical challenges were addressed in 2017, the need for efficient access and query on distributed semantically annotated datasets, and the need for efficient semantic annotation processes. Three methodological challenges were identified: semantic representation of large textual datasets, potential biases associated with semantic annotations, and computation of Bayesian indicators on data annotated with ontologies.

In one of the five best papers selected this year, Khan *et al.* [4] addressed the first technical challenge with SAFE, a SPARQL-federated query system for RDF data cubes with access control. There are some research works about SPARQL-federated query systems and access control for SPARQL query engines and the RDF Data Cube Vocabulary exists already as the world wide web consortium (W3C) recommendation for describing data cubes as RDF. However, the authors successfully integrated these three elements in SAFE and they compared its performance against existing SPARQL-federated query systems with clear advance on the latter. Besides the raw performances demonstrated in this work, there is no actual experiment in medicine that could prove its usefulness, although good results are expected. The second technical challenge is the semantic annotation and Cuzzola *et al.* [9] gave a good overview on existing approaches. At the frontier with the clinical Natural Language Processing (NLP) selection of papers for this edition of the Yearbook, the discussion and methods made this work also relevant to the KRM community. The semantic annotation is crucial for many KRM processes and the presented annotator (RysannMD) exhibited very promising results (precision, recall, and F1 measure and processing time) as it was demonstrated in a benchmarking experiment with other modern annotators (cTAKES, MetaMap, NOBLE Coder, and Neji). One key feature of this tool is an efficient semantic disambiguation that relies on the UMLS Semantic Network®. Additionally, this tool is immediately applicable.

Among the methodological challenges, Shi *et al.* [14] established a novel approach to integrate textual medical knowledge. With a specific model and NLP techniques, they converted medical texts into conceptual graphs and pruned meaningless inferences with an experimental algorithm. Although proper experiments on real datasets (electronic health records) were not presented in the paper, this approach represents a significant contribution for semantic processes over large medical corpora.

Kulmanov *et al.* [12] investigated the impact of the annotation size on a large number of measures of semantic similarity. This has become of major interest given the growing methods relying on similarity measures in particular in the field of omics data semantic analysis. They concluded that most measures were sensitive to the number of annotations per entities, to the difference in annotation sizes among compared entities, and to the

concepts' depth in the ontology. However, this work does not discuss the negative impact of these biases, neither present the methods to control them. In any case, further work on the potential solutions that could be used to quantify and control the effect of annotation size on similarity measures would be of major interest to the KRM community.

The last methodological challenge was addressed by Barton *et al.* [8], who provided an elegant theoretical basis for the use of ontologies for Bayesian indicators calculation, accounting for the granularity represented in these ontologies (i.e. the "spectrum effect"). This work introduced a meaningful method to derive the usual Bayesian indicators of performance (i.e. sensitivity, specificity, positive predictive value, and negative predictive value), which are mandatory indicators in the medical research, when data are annotated with ontologies.

Conclusions

After the refinement in ontology design and numerous initiatives for ontologies integration into knowledge-based systems observed in 2016, significant consequent results were published in 2017. The first major advances were in the field of genetics, where ontologies appeared fully instrumental to phenotype-genotype associations mining, mainly supported by semantic similarity measurements. The "routine" work in ontology and knowledge base design remains a significant part of KRM research, however with very high quality in methods.

The growing use of ontologies has led to identifying new challenges for 2018 KRM research: the query on distributed semantically annotated datasets, the efficiency of semantic annotation processes, the semantic representation of large textual datasets, the control of biases associated with semantic annotations, and the computation of Bayesian indicators on data annotated with ontologies.

Acknowledgements

We would like to thank Martina Hutter and Adrien Ugon for their support and the reviewers for their participation in the selection process of best papers for the IMIA Yearbook.

References

1. Dhombres F, Charlet J. Knowledge Representation and Management, It's Time to Integrate! Yearb Med Inform 2017;26(1):148-51.
2. Boudellioua I, Mahamad Razali RB, Kulmanov M, Hashish Y, Bajic VB, Goncalves-Serra E, et al. Semantic prioritization of novel causative genomic variants. PLoS Comput Biol 2017;13(4):e1005500.
3. Galeota E, Pelizzola M. Ontology-based annotations and semantic relations in large-scale (epi)genomics data. Brief Bioinform 2017;18(3):403-12.
4. Khan Y, Saleem M, Mehdi M, Hogan A, Mehmood Q, Rebholz-Schuhmann D, et al. SAFE: SPARQL Federation over RDF Data Cubes with Access Control. J Biomed Semantics 2017;8(1):5.
5. Notaro M, Schubach M, Robinson PN, Valentini G. Prediction of Human Phenotype Ontology terms by means of hierarchical ensemble methods. BMC Bioinformatics 2017;18(1):449.
6. Petegrosso R, Park S, Hwang TH, Kuang R. Transfer learning across ontologies for phenome-genome association prediction. Bioinformatics 2017;33(4):529-36.
7. Alonso-Calvo R, Paraiso-Medina S, Perez-Rey D, Alonso-Oset E, van Stiphout R, Yu S, et al. A semantic interoperability approach to support integration of gene expression and clinical data in breast cancer. Comput Biol Med 2017;87:179-86.
8. Barton A, Ethier JF, Duvauferrier R, Burgun A. An ontological analysis of medical Bayesian indicators of performance. J Biomed Semantics 2017;8(1):1.
9. Cuzzola J, Jovanovic J, Bagheri E RysannMD: A biomedical semantic annotator balancing speed and accuracy. J Biomed Inform 2017;71:91-109.
10. Esteban-Gil A, Fernandez-Breis JT, Boeker M. Analysis and visualization of disease courses in a semantically-enabled cancer registry. J Biomed Semantics 2017;8(1):46.
11. Gipson DS, Kirkendall ES, Gumbs-Petty B, Quinn T, Steen A, Hicks A, et al. Development of a Pediatric Adverse Events Terminology. Pediatrics 2017;139(1).
12. Kulmanov M, Hoehndorf R. Evaluating the effect of annotation size on measures of semantic similarity. J Biomed Semantics 2017;8(1):7.
13. Natale DA, Arighi CN, Blake JA, Bona J, Chen C, Chen SC, et al. Protein Ontology (PRO): enhancing and scaling up the representation of protein entities. Nucleic Acids Res 2017;45(D1):D339-D46.
14. Shi L, Li S, Yang X, Qi J, Pan G, Zhou B. Semantic Health Knowledge Graph: Semantic Integration of Heterogeneous Medical Knowledge and Services. Biomed Res Int 2017;2017:2858423.
15. Vita R, Overton JA, Sette A, Peters B. Better living through ontologies at the Immune Epitope Database. Database (Oxford) 2017;2017(1).
16. Zhang YF, Gou L, Zhou TS, Lin DN, Zheng J, Li Y., et al. An ontology-based approach to patient follow-up assessment for continuous and personalized chronic disease management. J Biomed Inform 2017;72:45-59.

Correspondence to:
Dr. Ferdinand Dhombres
Médecine Sorbonne Université, INSERM and APHP
Hôpital Universitaire Armand Trousseau
service de médecine fœtale
26 rue du Dr Arnold Netter
75012 Paris, France
E-mail: ferdinand.dhombres@inserm.fr

Appendix: Content Summaries of Selected Best Papers for the IMIA Yearbook 2018, Section Knowledge Representation and Management

Boudellioua I, Mahamad Razali RB, Kulmanov M, Hashish Y, Bajic VB, Goncalves-Serra E, Schoenmakers N, Gkoutos GV, Schofield PN, Hoehndorf R

Semantic prioritization of novel causative genomic variants

PLoS Comput Biol 2017;13(4):e1005500

This paper addresses the problem of how to distinguish which of the thousands of DNA sequence variants carried by an individual with a rare disease is (or are) responsible for the disease phenotype. This can help clinicians to reach a diagnosis, but also can be instrumental in improving the understanding of the underlying physiopathology of the disease. Many methods are currently available to help in identifying causative variant, *e.g.* by using information about species evolution and variant conservation or by the prediction of functional consequences of the variant DNA sequence. The authors has developed the PhenomeNET Variant Predictor (PVP) system that leverages semantic technologies and automated reasoning over genotype-phenotype relations to filter and prioritize variants in whole exome and whole genome sequencing datasets. In the heart of the system, a novel algorithm uses the patients' phenotype similarity to phenotypes in databases with known phenotype-genotype correlations to further rank potential candidate genes. In a retrospective study, they applied PVP to the interpretation of sequencing data in patients suffering from congenital hypothyroidism and showed that PVP accurately identified causative variants in whole exome and whole genome sequencing datasets. PVP provides a potentially significant resource for the discovery of causal variants.

The keys to the computational integration and comparison of phenotypes were (1) the systematic application of the PATO framework which provides a uniform way of describing phenotypes, and (2) the use of the UBERON ontology which can be used to systematically describe and relate anatomical structures between species. The PhenomeNET ontology also includes (as imports) GO (Gene Ontology), ZFA (Zebrafish Model Organism Database), CL (ontology for cell types), NBO (Neuro Behavior Ontology), ChEBI (database and ontology for chemical entities of biological interest), MPATH (the mouse pathology ontology), and others.

Galeota E, Pelizzola M

Ontology-based annotations and semantic relations in large-scale (epi)genomics data

Brief Bioinform 2017;18(3):403-12

This work investigates the possible use of public repositories for assembling data sets including chromatin immunoprecipitation assays with massive parallel sequencing (ChIP-seq) data that were widely under exploited.

The hypothesis is that using semantic annotation of the metadata of public data sets with concepts from biomedical ontologies allows for a common description and interoperability of these data. The authors demonstrated that these annotations efficiently support the retrieval of samples for a given condition of interest over several large repositories. Additionally, a process of clustering based on semantic similarity metrics resulted in large groups coherent samples. The comparison of tools based on the UMLS (Metamap on version 2014AA) with tools that use topic-specific OBO ontologies (Concept mapper on BRENDA Tissue Ontology (BTO) and Disease Ontology (DO)) showed that the latter outperforms the former both in the annotation process and in the computation of semantic similarity measures. But given the dates of the resources, it should be noted that the 2014AA version of UMLS did not include the HPO ontology.

This approach is positively assessed by a case-study on a set of semantically homogeneous ChIP-seq samples targeting a specific transcription factor (Myc) and expanded with semantically similar epigenetic samples. The semantic information proved to be coherent with the ChIP-seq signal and the current knowledge about this transcription factor.

Khan Y, Saleem M, Mehdi M, Hogan A, Mehmood Q, Rebholz-Schuhmann D, Sahay R

SAFE: SPARQL Federation over RDF Data Cubes with Access Control

J Biomed Semantics 2017;8(1):5

In this paper, the authors propose SAFE, a query federation engine that enables policy-aware access to sensitive statistical data sets represented as RDF data cubes. SAFE is designed specifically to query statistical RDF data cubes in a distributed setting, where access control is coupled with source selection, and user profiles and their access rights. SAFE proposes a join-aware source selection method that avoids wasteful requests to irrelevant and unauthorized data sources. In order to preserve anonymity and enforce stricter access control, SAFE's indexing system does not hold any data instances—it stores only predicates and endpoints.

SAFE is motivated by the needs of three clinical organizations in the context of a European Union (EU) project which aims at enabling controlled federation over statistical clinical data – such as data from clinical trials – owned and hosted in situ by multiple clinical sites, represented in the form of data cubes. However, the methods proposed by SAFE can be used in other settings involving data cubes outside of the Health Care and Life Sciences domain (even for open data).

The resulting data summary has a significantly lower index generation time and size compared to existing engines, which allows for faster updates when sources change. Moreover, the authors show that SAFE enables granular graph-level access control over distributed clinical RDF data cubes and efficiently reduces the source selection and overall query execution time when compared with general-purpose SPARQL query federation engines in the targeted setting.

Notaro M, Schubach M, Robinson PN, Valentini G

Prediction of Human Phenotype Ontology terms by means of hierarchical ensemble methods

BMC Bioinformatics 2017;18(1):449

The prediction of human gene–abnormal phenotype associations is a fundamental

step toward the discovery of novel genes associated with human disorders, especially when no gene is known to be associated with a specific disease. In this context, the Human Phenotype Ontology (HPO) provides a standard categorization of the abnormalities (phenotypes) associated with human diseases.

In this paper, the authors tackle the problem of learning associations when the annotation source is HPO, i.e. a formal ontology, per se. HPO, as a formal ontology, is structured as a direct acyclic graph, where more general classes are found at the top levels of the hierarchy and the class specificity increases moving towards the lower levels of the hierarchy, *i.e.* from root to leaves. As a consequence, each class may have more than one parent and such an ontology is governed by the annotation propagation rule: if a gene is annotated with a given functional class, then it is annotated with all the "parent" classes, and with all its ancestors in a recursive way. On the contrary if a gene is not annotated to a class, it cannot be annotated to its offspring.

The authors present two hierarchical ensemble methods that they formally prove to provide biologically consistent predictions according to the hierarchical structure of HPO. The modular structure of the proposed methods, that consists in a "flat" learning first step and a hierarchical combination of the predictions in the second step, allows the predictions of virtually any flat learning method to be enhanced. The experimental results show that hierarchical ensemble methods are able to predict novel associations between genes and abnormal phenotypes with results that are competitive with state-of-the-art algorithms and with a significant reduction of the computational complexity. The implementation of the proposed methods is available as an R package from the CRAN repository. This result is important because it enhances prediction algorithms (for gene–abnormal phenotype associations) when association annotations are described with formal ontologies.

Petegrosso R, Park S, Hwang TH, Kuang R Transfer learning across ontologies for phenome-genome association prediction Bioinformatics 2017;33(4):529-36

In this paper, the authors tackle the problem of learning associations when the Knowledge Organization System (KOS) of annotations is HPO, i.e. a formal ontology, per se. Moreover, they take into account the nature of the KOS describing genome, the Gene Ontology (GO). They note that there are only few known associations available for training. For example, in HPO, more than half of the phenotypes are annotated at best with only one gene association, and this sparsity makes prediction impossible or much less reliable even if gene–gene interactions can be introduced as additional training information.

The authors introduce Dual Label Propagation (DLP) to impose consistent associations with the entire phenotype paths in predicting phenotype–gene associations in HPO. DLP is then used as the base model in a transfer learning framework (tlDLP) to incorporate functional annotations in GO. By simultaneously reconstructing GO, term–gene associations and HPO phenotype–gene associations for all the genes in a protein–protein interaction network, tlDLP benefits from the enriched training associations indirectly through relations with GO terms.

In their experiments to predict the associations between human genes and phenotypes in HPO based on human protein–protein interaction network, both DLP and tlDLP improved the prediction of gene associations with phenotype paths in HPO in cross-validation and the prediction of the most recent associations added after the snapshot of the training data. Moreover, the transfer learning through GO term–gene associations significantly improved association predictions for the phenotypes with no more specific known associations by a large margin.

Finally, the paper suggests that transfer learning can fulfill prediction with missing training information by the relation among GO and HPO. The results on predicting GO gene functions further support the conclusion that transfer learning across the two domains is beneficial to both learning tasks.